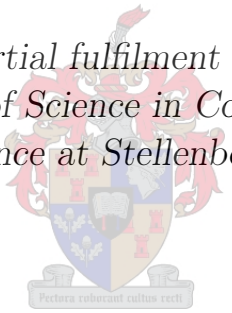
Analysing Ranking Algorithms and Publication Trends on Scholarly Citation Networks

by

Marcel Dunaiski

Thesis presented in partial fulfilment of the requirements for the degree of Master of Science in Computer Science in the Faculty of Science at Stellenbosch University



Computer Science Division,
Department Mathematical Sciences,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Supervisors:

W. Visser J. Geldenhuys

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: 2014/08/31 .....

Abstract

Analysing Ranking Algorithms and Publication Trends on Scholarly Citation Networks

M.P. Dunaiski

Computer Science Division,
Department Mathematical Sciences,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: MSc

August 2014

Citation analysis is an important tool in the academic community. It can aid universities, funding bodies, and individual researchers to evaluate scientific work and direct resources appropriately. With the rapid growth of the scientific enterprise and the increase of online libraries that include citation analysis tools, the need for a systematic evaluation of these tools becomes more important.

The research presented in this study deals with scientific research output, i.e., articles and citations, and how they can be used in bibliometrics to measure academic success. More specifically, this research analyses algorithms that rank academic entities such as articles, authors and journals to address the question of how well these algorithms can identify important and high-impact entities.

A consistent mathematical formulation is developed on the basis of a categorisation of bibliometric measures such as the h-index, the Impact Factor for journals, and ranking algorithms based on Google's PageRank. Furthermore, the theoretical properties of each algorithm are laid out.

The ranking algorithms and bibliometric methods are computed on the Microsoft Academic Search citation database which contains 40 million papers and over 260 million citations that span across multiple academic disciplines.

We evaluate the ranking algorithms by using a large test data set of papers and authors that won renowned prizes at numerous Computer Science conferences. The results show that using citation counts is, in general, the best ranking metric. However, for certain tasks, such as ranking important papers or identifying high-impact authors, algorithms based on PageRank perform better. As a secondary outcome of this research, publication trends across academic disciplines are analysed to show changes in publication behaviour over time and differences in publication patterns between disciplines.

Opsomming

Analise van rangalgoritmes en publikasie tendense op wetenskaplike sitasienetwerke

M.P. Dunaiski

Rekenaarwetenskap Afdeling,
Departement van Wiskundige Wetenskappe,
Universiteit van Stellenbosch,
Privaatsak X1, Matieland 7602, Suid Afrika.

Tesis: MSc

Augustus 2014

Sitasiesanalise is 'n belangrike instrument in die akademiese omgewing. Dit kan universiteite, befondsingsliggams en individuele navorsers help om wetenskaplike werk te evalueer en hulpbronne toepaslik toe te ken. Met die vinnige groei van wetenskaplike uitsette en die toename in aanlynbiblioteke wat sitasieanalise insluit, word die behoefte aan 'n sistematiese evaluering van hierdie gereedskap al hoe belangriker.

Die navorsing in hierdie studie handel oor die uitsette van wetenskaplike navorsing, dit wil sê, artikels en sitasies, en hoe hulle gebruik kan word in bibliometriese studies om akademiese sukses te meet. Om meer spesifiek te wees, hierdie navorsing analiseer algoritmes wat akademiese entiteite soos artikels, outeers en journale gradeer. Dit wys hoe doeltreffend hierdie algoritmes belangrike en hoë-impak entiteite kan identifiseer.

'n Breedvoerige wiskundige formulering word ontwikkel uit 'n versameling van bibliometriese metodes soos byvoorbeeld die h-indeks, die Impak Faktor vir journaale en die rang-algoritmes gebaseer op Google se PageRank. Verder word die teoretiese eienskappe van elke algoritme uitgelê.

Die rang-algoritmes en bibliometriese metodes gebruik die sitasiedatabasis van Microsoft Academic Search vir berekeninge. Dit bevat 40 miljoen artikels en meer as 260 miljoen sitasies, wat oor verskeie akademiese dissiplines strek.

Ons gebruik 'n groot stel toetsdata van dokumente en outeers wat bekende pryse op talle rekenaarwetenskaplike konferensies gewen het om die rang-algoritmes te evalueer. Die resultate toon dat die gebruik van sitasietellings, in die algemeen, die beste rangmetode is. Vir sekere take, soos die gradeering van belangrike artikels, of die identifisering van hoë-impak outeers, presteer algoritmes wat op PageRank gebaseer is egter beter. 'n Sekondêre resultaat van hierdie navorsing is die ontleding van publikasie tendense in verskeie akademiese dissiplines om sodoende veranderinge in publikasie gedrag oor tyd aan te toon en ook die verskille in publikasie patrone uit verskillende dissiplines uit te wys.

Contents

List of Figures

List of Tables

Nomenclature

Acronyms The acronyms listed here are not exhaustive. Acronyms used to refer to various conferences and publishing sources are listed separately in Appendix A.

AF	Author-Level Eigenfactor
AI	Article Influence
AP	average precision
CC	Citation Count
CCR	(Citation-)CountRank
CR	CiteRank
CS	Computer Science
CW	census window
EF	Eigenfactor
ICSE	International Conference on Software Engineering.
IF	(Journal) Impact Factor
MAP	mean average precision
MAS	Microsoft Academic Search
MIP	most influential paper
NR	NewRank
PC	Publication Count
PR	PageRank
PRA	PageRank for Authors
PRV	PageRank for Venues
SR	SceasRank
SR1	SceasRank1
SR2	SceasRank2
TW	target window
YR	YetRank

Variables

r	Pearson's correlation coefficient
ρ	Spearman's rank correlation coefficient
τ	Kendall's Tau-b rank correlation coefficient; characteristic time decay variable
p	paper
i, j, k	indices for elements in a set

$\mathcal{P}$	set of papers
$\mathcal{A}$	set of authors
$\mathcal{V}$	set of venues
$\mathcal{Y}$	set of publication years of papers
t	iteration
α	damping factor
δ	precision threshold
$\boldsymbol{x}$	result vector

Graph Notation

G	bibliometric citation network or general graph
u, v	vertices of a graph
e	edge of a graph
$V(G)$	vertex set of graph G
$E(G)$	edge set of graph G
N, n	the order of a graph. For example, $n = \|V(G)\|$ is the order of graph G
m	the size of a graph, $m = \|E(G)\|$
$N_G^+(v)$	out-neighbourhood of vertex v in graph G
$N_G^-(v)$	in-neighbourhood of vertex v in graph G
$\mathrm{od}_G(v)$	out-degree of vertex v in graph G
$\mathrm{id}_G(v)$	in-degree of vertex v in graph G
w_{ij}	weight associated with the edge from vertex i to vertex j

Chapter 1

Introduction

Counting citations as an evaluation metric for academic journals was first proposed in 1927 by two chemists, Gross and Gross, at Pomona College in California [1]. Due to the increasing size and specialization of academic fields, they saw the need for small libraries with limited financial resources to methodically rank journals in order to decide which periodicals to subscribe to.

Since then a lot of research has been conducted on how to best measure the value of scientific entities such as papers, authors, journals and universities. This is now known as bibliometric citation analysis and is an important aspect of the scientific knowledge process with many applications. For instance, it assists researchers in deciding where to publish their work, aids funding bodies in distributing financial resources, and helps university review panels to evaluate tenure candidates.

The most prominent and widely used metrics today are the Impact Factor for journals and the h-index for authors. The Impact Factor was first introduced by Garfield [2] in 1955 and ranks journals according to the average number of citations that they receive in two years. The h-index, proposed by Hirsch [3] in 2005, is also based on citation counts and the number of papers that an author has published.

The research presented in this thesis deals with these bibliometric measures and various ranking algorithms that are based on Google's PageRank [4] and that can be adapted for scientific citation networks. These metrics are categorised and defined in a concise mathematical formulation. The focus of this research is on the comparison of these ranking algorithms and their evaluation using large test data sets that are based on expert opinions.

Two academic citations data sets are used to construct citation networks. Firstly, the Microsoft Academic Search (MAS) data set is used for all of the experiments in this paper [5]. This data set contains 40 million papers and over 260 million citations spanning across different academic disciplines. Secondly, a data set obtained from Tang *et al.* [6] is used comparatively which is based on the DBLP database [7] and comprises Computer Science papers.

1.1 Motivation

In the last few decades the research community has seen a rapid growth in the output of academic publications. Ever more academic work is published electronically and accurate meta-information about publications is becoming more available.

This has important implications. Firstly, it changes the way in which academics conduct their research. They have easier access to more information and cite more online sources. With this, the speed at which scientific output is produced has increased. Secondly, it changes how and by whom scientific products are evaluated. For example, institutions such as universities have increasing access to real-time instruments and can apply a variety of metrics to evaluate researchers. It also creates more opportunities to better evaluate these metrics. More importantly, it opens the possibility of analysing the meta-data to discover previously unknown properties of the publication processes.

The task of searching and indexing information is moving away from librarians towards software. Computers are very good at indexing machine-readable information and handling search queries by returning results that fulfill the user's query. However, it is much more difficult for a computer to reason about the quality of information in order to decide, for example, which paper is the most relevant in its field. Therefore, it has become increasingly important to devise adequate ranking and evaluation tools to help researchers find exactly the information they need.

Nowadays, the most widely adopted metrics used to judge a paper's importance are based on citation counts. Using citation counts is an easy and intuitive metric for calculating a paper's importance, but it has certain drawbacks and limitations. The problem with merely counting a paper's citations is that the results can be skewed and do not necessarily represent the real value of a paper [8].

Currently, the most widely adopted metric for judging a journal's impact is the Journal Impact Factor [2]. The main critique of the Journal Impact Factor is that it varies between disciplines and depends on the speed at which papers get cited. Furthermore, the Impact Factor only calculates the overall impact of a journal and does not measure the influence of the papers published in a journal.

The Eigenfactor metric devised by Bergstrom *et al.* [9] tries to overcome the drawbacks of the Journal Impact Factor. It is based on the PageRank algorithm and computes overall impact scores for journals and a per article influence for the papers published at journals.

The h-index [3] was originally devised to compare the impact of researchers but can be adapted to measure the influence of journals and universities as well. The main disadvantage of the h-index is that it can only be used to compare the impact of authors that are in roughly the same stages of their careers and that it cannot be used to compare authors that work in different academic fields.

The Author-Level Eigenfactor [10], which uses the same approach as the Eigenfactor metric for journals, computes influence scores for authors by adapting the PageRank algorithm and using a co-author citation graph as input.

A lot of research has been conducted on paper ranking algorithms to identify important papers. Most approaches use a PageRank-like algorithm with various alterations such as incorporating the publication dates of papers or the impact factors of journals into their computations [11; 12; 13; 14].

The above-mentioned metrics have different approaches and applications. For instance, the Author-Level Eigenfactor algorithm [10] can only be used to rank authors while algorithms such as CiteRank [13] and SceasRank [11] are only applicable to papers.

The problem is that all these various algorithms have not been compared and evaluated extensively. Some metrics, such as the SceasRank algorithm, have been compared to the basic PageRank algorithm and evaluated using a small test data set [11]. Similarly, Dunaiski and Visser [15] compare some algorithms and evaluate them using a small set of papers that won prizes for their influence at a single conference. Nonetheless, comprehensive comparisons and evaluations of these algorithms have not been researched sufficiently.

The research presented in this thesis fills this knowledge gap by classifying and comparing the various algorithms and evaluating them using a large test data set that is based on expert opinions.

For the evaluation of the algorithms, four large test data sets were compiled that are used for four different evaluation purposes.

Firstly, a data set that contains papers that won prizes for their high impact in their fields was collected. These prizes are awarded about 10 years after their initial publication in recognition of their influence in the last decade. These papers are used to evaluate how well the ranking algorithms can identify high-impact papers.

Secondly, a data set of authors that have won prizes for their outstanding, innovative and long-lasting contributions to their fields was compiled. This test data is used to evaluate how well the author ranking algorithms identify important researchers.

Thirdly, a list of papers that won best-paper awards at different conferences was compiled. Conference committees or Special Interest Groups of organisations award best-paper prizes at conferences to papers that were selected by a review panel in the year of publication. This set of papers is used to assess how well the review panels of the various venues can predict high-impact papers.

Lastly, a list of papers that had a high influence in Computer Science was compiled. Using this data set, the paper ranking algorithms are evaluated on how well they can identify overall important papers.

The research presented in this paper provides further insight into the ranking of academic entities, with a focus on paper ranking algorithms. The algorithms are compared empirically by looking at their computed rankings. The goal is to identify properties of the ranking algorithms that influence the way they rank papers, authors and venues that can be used for the development of new bibliometric measures.

1.2 Research Questions and Objectives

The problem with all algorithms and metrics that are based on citation counts is the interpretation of what a citation means and how citations should be weighted to compute fair ranking scores. Should a citation from a renowned journal be weighted more because of its status? Or should it be weighted less so as not to overshadow small but still significant journals? Moreover, how can we account for the fact that recently published papers have not been around very long and therefore have not accrued a lot of citations? Should the age of a publication be considered when computing rankings? Furthermore, should citation ages be taken into account? After all, the direct citation of an older paper by a newer paper might indicate that it still bears current relevance. Should citations be weighted depending on academic fields? Different fields have different citation conventions and might impact the results of ranking algorithms.

A discussion of citation counts and their use is a sensitive and controversial issue [8]. A clear distinction has to be made between impact, significance and quality [16, p. 7].

How should self-citations be counted when computing importance scores? Some believe that self-citations manipulate citation rates, while others believe that it is very reasonable since it is an indication of a narrow speciality where scientists tend to build on their own work and that of collaborators [8].

Furthermore, can papers that stopped receiving citations due to obliteration be identified even though they are still of importance but their work is so ingrained in the body of knowledge that they are not cited anymore? How can significant papers be identified that are far ahead of their field and go unnoticed until the field catches up?

The above-mentioned questions have to be considered when designing fair ranking algorithms for papers, authors and venues. Some of these problems have been worked on recently [13; 14; 11; 15], but a concise analysis of the properties of the ranking algorithms has not been conducted. Furthermore, an in-depth comparison and evaluation of more than a small subset of algorithms has also not been performed.

From the points outlined above, the following objectives have been identified and are pursued in this thesis:

- Research bibliometric measures to obtain a deeper understanding of ranking algorithms that can be used to rank academic entities.
- Define the various ranking algorithms uniformly using a consistent notation for better comparability.
- Obtain further knowledge about the publication processes and trends that occur in the production of scientific output.
- Collect a large test data set that can be used to evaluate the ranking algorithms.
- Identify the properties of the various ranking algorithms and find the best suited algorithm for identifying important and high-impact papers and influential authors.

1.3 Thesis Overview

Chapter 2 provides background information about the field of bibliometrics and how citation analysis can be used to rank articles and journals. Background information about Markov chains and the PageRank algorithm is also given. In addition, this chapter shows how these can be adapted to citation networks to rank papers.

Chapter 3 begins by outlining the history of bibliometrics and scientometrics, followed by an in-depth discussion on what impact, quality and importance of papers mean and what citation counts can measure. In addition, a literature review of current algorithmic approaches to rank papers, authors and publication venues is given.

Chapter 4 contains detailed descriptions of ranking metrics that are based on pure citation counts and algorithmic approaches to ranking academic entities. Ranking algorithms are defined mathematically using uniform and concise formulations. The theoretical advantages and drawbacks of each ranking algorithm are discussed.

Chapter 5 details the citation data sets used in this thesis and the test data that was collected for this research. Some publication trends are discussed and how they differ between academic disciplines.

Chapter 6 compares the paper-, venue- and author-ranking algorithms empirically by analysing their ranking outputs directly to identify ranking properties.

Chapter 7 shows the results of evaluating the paper and author algorithms with test data that is based on expert opinions.

Chapter 8 concludes the research by briefly reiterating and discussing the main results that were obtained and describes possible future research avenues related to this thesis.

Chapter 2

Background Information

2.1 From Bibliometrics to Cybermetrics

It can be quite difficult to classify the research presented in this thesis and to assign it to specific and well-known research fields. It touches upon several topics that may appear unrelated and are not usually discussed together. Furthermore, as is often the case, there is no general consensus with regards to the formal definition of many terms.

In a broad sense, the research falls under the umbrella field of information science and touches upon four narrower research fields, namely: scientometrics, informetrics, cybermetrics and, in particular, bibliometrics. These fields, as described by Hood and Wilson [17, p. 1], are component fields related to the study of the dynamics of disciplines as reflected in the production of their respective bodies of literature.

To define these fields more narrowly, one has to look at where their names first appeared and in which contexts they are used.

The field of *scientometrics* can be closely linked to two people. Vassily Nalimov coined the equivalent Russian term "Naukometriya" in 1973 [18, p. 2], and T. Braun translated the term for the journal "Scientometrics" which was founded in 1977 [19, p. 1]. Since then the term scientometrics has gained popularity and is used to describe research that is committed to the study of the growth, structure, interrelationships, and productivity of science [17, p. 1].

According to Hood and Wilson [17, p. 3], a lot of scientometrics is indistinguishable from bibliometrics and a sizeable amount of bibliometric research is published in the journal "Scientometrics". The differentiator between the two fields is that bibliometrics focuses only on the literature output of science. Scientometrics, on the other hand, is a more general field and incorporates more aspects of science than merely its literature. For example, scientometrics is also concerned with the practices of research, research and development management, and the study of law related to science and technology.

According to Tague-Sutcliffe [19, p. 1], *bibliometrics* focuses on quantitative studies surrounding the creation, spreading, and recording of scientific information by developing mathematical models to help in the prediction and decision-making of the scientific enterprise. Therefore, the research presented in this document most closely fits into the field of bibliometrics, since it focuses on data created by the literature output of the scientific community and how this information can be analysed in order to gain additional knowledge about the sciences. Of course, there may also be a political dimension to the use of citation analysis, but any discussion of this dimension would go beyond the scope of this thesis.

Informetrics is the most general term and subsumes scientometrics and bibliometrics. Tague-Sutcliffe [19, p. 2] describes informetrics as the study of literature, documents, and the mathematical properties of the laws and distribution of information. Hence, informetrics does not focus only on scientific publications and bibliographies, but any type of measurable information such as metrics of the Internet, social networks, or the dissemination of public information.

Lastly, the term *cybermetrics* should also be mentioned here. The journal "Cybermetrics" covers the fields of scientometrics, informetrics and bibliometrics, with a focus on their interrelationship with the Internet [20].

The research presented here tangentially touches on the field of cybermetrics since the data sets that are used for the citation networks are a result of the Internet and how research is currently conducted.

2.2 Notation and Terminology

As far as possible, consistent notation is used throughout this document to reduce confusion or misinterpretation of information. In cases where this is more difficult, the context will provide the reader with the necessary information to understand each symbol, or it will be explained immediately afterwards.

In this document, the terms *article* and *paper* are used indistinguishably and refer to some written work that is in some stage of the publishing process. It is a very generic definition that encompasses any written text, from Masters theses to scientific short communications and books, and can be pre-print versions, published articles or re-published papers.

A venue refers to the place of publication and usually has a one-to-many relationship with authors and papers. Most commonly, a venue refers to a journal that contains a number of articles or to a conference where a set of papers are published. The term venue can also define a broader concept of a collection of papers or authors. For example, academic departments at a university, research institutes, or commercial entities could be viewed as publication venues that publish work which is incorporated into the general body of academic knowledge.

The affiliation of an author is the place of work associated with a published paper, at the time of publication. Science can be divided into several domains and subdomains. On the one hand, this division is largely subjective, but on the other hand, it is important because writing style and citation culture differ and have an impact on the results of citation analysis. This is further discussed in Section 5.4 where citation analysis is performed on data that is divided into different domains.

The symbol p is used to refer to papers and may be subscripted, such as p_i or p_j, if more than one paper is referenced. Similarly, a, v, y refer to authors, venues and years, respectively. The symbols $\mathcal{P}$, $\mathcal{A}$, $\mathcal{V}$ and $\mathcal{Y}$ indicate, respectively, sets of papers, authors, venues and years.

Bold characters, such as $\boldsymbol{x}$ and $\boldsymbol{\rho}$, are used to represent vectors, as are acronyms of ranking algorithms. For example, $\boldsymbol{PR}$ is a result vector that contains PageRank values.

Two different norms for vectors are used in this thesis; the L^1-norm and the L^2-norm. The L^2-norm is the commonly known Euclidean norm and is indicated by $\|\boldsymbol{x}\|$. Instead of the Euclidean distance, the grid distance or "Manhattan distance" can be used for the norm of vectors: this L^1-norm is explicitly indicated with a subscripted 1 and defined as $\|\boldsymbol{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|$.

2.2.1 Graph Notation

Citation networks are directed graphs where papers are vertices and citations are edges connecting two vertices in the graph. Let G be a directed graph of order n and size m, where $V(G)$ is the vertex set containing n vertices and $E(G)$ is the edge set of size m. The shorthand notation $G = (V, E)$ is sometimes used to describe a graph with a vertex set V and edge set E. Two vertices $u, v \in V$ are adjacent if the edge $e = (u, v) \in E$. For citation networks the directed edge $e = (u, v)$ implies that paper u references paper v.

The degree of a vertex v is the number of edges connected to v, denoted by $d(v)$. In a directed graph G, the out-degree of a vertex v, denoted $\text{od}(v)$, is the number of edges that start at v, and the in-degree, denoted $\text{id}(v)$, is the number of edges that terminate at v. Therefore, $d(v) = \text{id}(v) + \text{od}(v)$.

The adjacency matrix of the graph G, denoted by A, is an $n \times n$ binary matrix whose (i, j)-th element is 1 if $(v_i, v_j) \in E(G)$ and 0 otherwise.

In a weighted directed graph, edges may have weights associated with them. In this document, edge weights are assumed to be non-negative real values denoted by $w(e)$ where $e = (u, v)$ is the edge from vertex u to vertex v. The shorthand notation w_{uv} is used.

The out-neighbourhood of a vertex v in a directed graph G is the set $N_G^+(v) = \{u \in V(G) | (v, u) \in E(G)\}$. Similarly, the in-neighbourhood of a vertex v in a directed graph G is the set $N_G^-(v) = \{u \in V(G) | (u, v) \in E(G)\}$. It follows that $\text{od}_G(v) = \left|N_G^+(v)\right|$, while $\text{id}_G(v) = \left|N_G^-(v)\right|$.

The above notation can be used to describe a citation network. Let G be a directed graph representing the network of n papers and m citations. Then $V(G)$ is the set of papers and $E(G)$ is the set of citations. Furthermore, let $p_i, p_j \in V(G)$, then paper p_i references paper p_j if edge $(p_i, p_j) \in E(G)$. Using this graph notation to describe a paper p, the references in paper p's reference list is the set $N_G^+(p)$, while the number of citations that paper p received is $\text{id}(p)$.

2.3 Using Citations for Ranking

Citation is a research concept and a fundamental idea behind science. It facilitates collaboration, the re-use of previous work and the advancement of science as a whole. More specifically, and in the context of citation networks, citations are the predominant method to acknowledge the use of someone else's ideas, add credibility and verifiability to your own work, and to avoid plagiarism.

The physical manifestation of citations are often a list of references to other work in a bibliography section at the end of an article. There are different citation and referencing styles in the academic community and therefore it is important to clearly define the concept of citations and references. A *citation* generally refers to an acknowledgement of other work within the body of text. A *reference*, in turn, is the corresponding detailed literature reference included in the bibliography or literature cited section of published work and is normally found at the end of the text.

In the context of citation networks and for the purpose of this document, the terms reference and citation are used to distinguish between outgoing references from a paper and incoming citations to a paper. The terms are therefore used slightly differently from the traditional sense and are defined as follows:

- A paper's *references* are the set of papers that it cites and that are included in the reference list of the current paper. In graph terminology, the equivalent of a paper's

references is the out-neighbourhood of the paper. Therefore, the out-degree of the vertex corresponding with a paper is the size of the paper's reference list.

- A paper's *citations* are the set of papers, published after the current paper was made available, which cite the paper. Therefore, the vertices associated with this set of papers constitute the in-neighbourhood of the current paper in the citation network. Accordingly, the in-degree of a vertex is the number of times the corresponding paper has been cited since its publication.

It should be noted that not all papers contain references or citations. A paper may not have citations associated with it because it has not been cited or because citations are not identified correctly and therefore are not included in a data set. Some referencing styles use in-text citations only and do not contain a bibliography section at the end of an article. Often, these references are not indexed in bibliographic databases. Ultimately, the completeness and correctness of the references used in a citation network depend on the data source that is used and the reference mining method that is applied to extract references from papers.

Citations, as defined above, can be used to measure the impact of the work that is being cited, since by its nature it is some kind of acknowledgement. *Citation analysis* is based on this observation and by simply counting citations, various impact metrics can be defined:

- The total citation count of an article can be used as an indicator of its importance.
- The total or average citation counts of an author's papers can be used as an indicator of the impact of the author's contribution to the scientific corpus.
- The average citation counts of articles published in a journal can indicate the importance of the journal within its domain.

Citation analysis is also used to group similar papers together into clusters for recommending papers to researchers. Two early methods in citation analysis are bibliographic coupling [21] and co-citation [22], both of which identify closely-related papers. These methods are based on the idea that related publications share identical references or are cited by the same papers. In co-citation, the more citations two papers have in common the more closely they are related. Similarly, in bibliographic coupling, the more papers that are listed in both papers' reference lists, the more closely related they are considered to be.

Citation analysis is also used in methods that fall into a category that can be classified under the broad term of *Journal Ranking*. These methods can be used to rank venues such as journals, conferences, academic departments, and authors. In other words, given an entity that publishes one or more papers, these methods can be used to compute an associated score. The interpretation of this computed score depends on the proposed purpose of the method that was used. In general, the results of journal ranking methods are intended to reflect the importance of journals within their field, the relative difficulty of publishing in a specific journal, and the prestige associated with publishing in a certain journal. Examples of methods that can be used to rank journals are the h-index and the Journal Impact Factor which are explained in more detail in Sections 4.1 and 4.3, respectively.

2.4 Markov Chains

Citation networks as a whole can also be taken into account and used as a basis to compute rankings for individual papers instead of simply counting a paper's number of citations. This section presents background information on Markov chains and how ranking algorithms make use of a citation network's entire structure to compute ranks. In the following sections, an analogy of a *random researcher* traversing the citation network is used each time a mathematical property is introduced or defined, in order to give an intuitive description of how these ranking algorithms use Markov chains to rank papers.

When using a Markov chain model, each paper in a citation network is regarded as a state. A citation from one paper to another is considered a transition which leads from one state to another state with a certain probability. Intuitively, this models a random researcher arbitrarily following a citation in a paper's reference list as a state transition in the Markov chain.

The idea behind using Markov models on citation networks is that if certain properties of the model (which are discussed in this section) hold true it is possible to compute the *steady-state distribution* of a Markov chain [23, ch. 17]. In the context of random researchers, this steady-state distribution represents the average proportion of time spent at a vertex in the citation network, which in turn can be interpreted as the importance of the corresponding paper because it signifies the interest of random researchers in the paper.

Let $\mathbf{X}_t$ be the state of the Markov chain at time t. $\mathbf{X}_t$ is not known with certainty before time t and may be viewed as a random variable. The description of the relation between the random variables $\mathbf{X}_0$, $\mathbf{X}_1$, $\mathbf{X}_2$, ... is called a *discrete-time stochastic process*. A random walk on a citation network is a discrete-time stochastic process since the position of the random researcher can only be observed at intervals, each time the random researcher follows a citation to another state.

Definition 1. A discrete-time stochastic process is a *Markov chain* if, for $t = 0, 1, 2, \ldots$ and all states i_t

$$\begin{aligned} &P(\mathbf{X}_{t+1} = i_{t+1} | \mathbf{X}_t = i_t, \mathbf{X}_{t-1} = i_{t-1}, \ldots, \mathbf{X}_1 = i_1, \mathbf{X}_0 = i_0) \\ &\quad = P\left(\mathbf{X}_{t+1} = i_{t+1} | \mathbf{X}_t = i_t\right) \end{aligned}$$

The definition says that the probability distribution of the state at time $t+1$ depends on the state i_t at time t, and does not depend on the states the chain passed through on the way to i_t. For random researchers, this definition implies that their choice of which citation to follow only depends on the entries of the current paper's reference list and not on the previously read articles that led the researcher to the current article.

Furthermore, we assume that for all states i and j and all t, the probability $P(\mathbf{X}_{t+1} = j | \mathbf{X}_t = i)$ is independent of t. This assumption allows us to write

$$P(\mathbf{X}_{t+1} = j | \mathbf{X}_t = i) = p_{ij} \tag{2.4.1}$$

where p_{ij} is the probability that given the system in state i at time t, it will *transition* to a state j at time $t+1$.

The p_{ij}'s are the corresponding *transition probabilities* for the Markov chain. Equation 2.4.1 implies that the probability law relating to a transition from the current state to the next state does not change over time (or is independent of t). A Markov chain that satisfies this equation is called a *stationary Markov chain*. Random researchers that

choose random citations to follow can be modeled as stationary Markov chains if a random researcher's choice of which citation to follow does not depend on how many citations he or she has followed before reaching the current paper.

For a Markov chain, the *initial probability distribution* is the vector $\boldsymbol{q}$ containing the probabilities of the chain for all states i at time 0. More formally the value $P(\mathbf{X}_0 = i) = q_i$ denotes the probability of the process of starting in state i. Therefore, the initial probability distribution $\boldsymbol{q}$ contains the probabilities of random researchers starting their search at certain papers.

Assuming that the state at time t is i, the process must be somewhere at time $t+1$. This means that for each state i,

$$\sum_{j=1}^{N} P\left(\mathbf{X}_{t+1} = j | P(\mathbf{X}_t = i)\right) = 1$$
$$\sum_{j=1}^{N} p_{ij} = 1 \tag{2.4.2}$$

P is the transition probability matrix of the Markov chain and if it satisfies Equation 2.4.2 then it is called a stochastic transition matrix. For a random researcher traversing a citation network this means that, with a probability of 1, he or she has to choose a reference to some vertex in the graph and cannot stay idle in the same state[1].

Therefore, using a transition matrix to describe citation networks, implies that each vertex in the network has at least one outgoing edge to another vertex in the network. This is not true for citation networks since papers exist that do not contain any references to other papers or data sets are incomplete and therefore contain papers with references that point outside the scope of the data set. Transition matrices have to be altered to meet this requirement so that one may use Markov chains to model random walks on citation networks. Some possible approaches of modifying citation networks to be stochastic are described in Section 2.4.1.

In order to compute a discrete result that captures the probabilities of the random researchers reaching certain vertices in the citation network, additional requirements are placed upon Markov chains.

A state is recurrent if the system will return to it with a probability of 1. For citation networks, this implies that in order for a state to be recurrent, a random researcher has to be able to return to a current paper by following citations. This does not hold true for citation networks. In Section 2.4.1 different ways of modifying the citation network are described so that this requirement is satisfied.

Definition 2. A state i is *periodic* with period $k > 1$ if k is the smallest number such that all paths leading from state i back to state i have a length that is a multiple of k. If a recurrent state is not periodic, then it is *aperiodic*.

Two states communicate with each other if they are accessible from each other.

Definition 3. If all states in a chain are recurrent, aperiodic, and communicate with each other, the chain is said to be *ergodic*.

[1]It is possible for a random researcher to stay with the same paper while keeping the transition matrix stochastic if a paper contains a single citation to itself with a corresponding probability of 1. This, however, does not fulfill the requirement of an ergodic chain (see Definition 3 in this section), since this vertex does not communicate with any other vertices in the graph.

Theorem 1. *Let P be the transition matrix for an N-state ergodic Markov chain. Then there exists a vector $\pi = [\pi_1 \quad \pi_2 \quad \cdots \quad \pi_N]$ such that*

$$\lim_{n\to\infty} P^n = \begin{bmatrix} \pi_1 & \pi_2 & \cdots & \pi_N \\ \pi_1 & \pi_2 & \cdots & \pi_N \\ \vdots & \vdots & & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_N \end{bmatrix}$$

The resulting vector π is called the *steady-state distribution* of the Markov chain and contains the average proportion of time that random researchers spend at specific vertices in the citation network.

In summary, a stationary Markov chain with a transition matrix P has a unique *steady-state distribution* if all states communicate with each other and are *aperiodic*. Citation networks are inherently non-ergodic since vertices are not recurrent because of the fact that papers can only reference older papers that have already been published and reference lists cannot be updated once articles have been published. Therefore, citation networks are intrinsically acyclic[2]. This implies that states in a Markov chain that is used to model random walks on a citation network, cannot be recurrent and hence also do not communicate with each other. In the following section, ways of modifying citation networks to obtain ergodicity are discussed.

2.4.1 Modelling Citation Networks using Markov Chains

To guarantee the existence of a steady-state distribution of a Markov chain, it has to be ergodic. The transition matrix of a citation network is inherently non-ergodic and therefore the chain's transition matrix has to be adapted to ensure that all states in the chain are recurrent, aperiodic and communicate with each other.

There are several ways to achieve this for citation networks with varying implications. The underlying graph and the impact on the precision of the results should be considered when choosing a method.

1. For each paper that does not contain any references (dangling vertex) to other papers, an edge is inserted from that paper's vertex to another random vertex within the graph. This alteration to the transition matrix heavily influences the steady-state distribution since the entire weight of a dangling vertex is transferred to a single vertex that is randomly selected.

2. Add N edges from each dangling vertex to all other vertices within the graph, including the dangling vertex itself. The weight is evenly distributed between the added edges such that each edge has a weight of $\frac{1}{N}$. This approach is the most accurate but increases the size of the graph substantially. This method for modelling citation networks for PageRank-like algorithms is used for ranking algorithms discussed in this thesis.

3. Remove all vertices from the graph corresponding to papers which contain no references. Since these vertices do not have any outgoing edges they do not influence

[2]It is assumed that for the sake of this argument reference lists of pre-printed articles are not updated. Theoretically, it is possible to create a citation cycle by referencing a pre-print article which adds a reference back to the citing article before final publication. In general this is not the case and since every vertex in the citation network has to be recurrent for ergodicity this assumption seems safe.

the value of the other vertices within the graph directly. After the steady-state distribution is computed for the other vertices, the scores for the dangling vertices can easily be calculated by reintroducing them into the graph. The disadvantage of this approach is that the transition probabilities from the vertices that remained in the graph, but had edges removed due to the pruning of the dangling vertices, will be affected. The advantage of pruning all dangling vertices is that it reduces both the order and the size of the graph, which can be substantial if the number of dangling vertices in a graph is large and therefore decrease the computation times considerably.

4. All vertices associated with papers that contain no references are combined into a single vertex. Lee *et al.* [24] show that combining dangling vertices in a Markov chain associated with PageRank and computing their results separately decreases the computation time of PageRank significantly. After computing the results for the dangling and non-dangling subsets of a graph, the results can be merged to obtain accurate approximations of the PageRank results.

2.5 The PageRank Algorithm

In this section, background information on the PageRank algorithm is provided since most algorithms used in this document are variations of the PageRank algorithm and are based on the same principles. The details described here focus on the mathematics and the computation of the algorithm with respect to citation networks. This section can be skipped if the reader's interest lies in the application of PageRank to academic citation networks. The PageRank algorithm for citation networks is defined separately in Section 4.2.1.

Essentially, the PageRank algorithm models a random walk on the citation graph of the Internet and, by means of the power method described below, computes the steady-state distribution of the Markov chain.

Let u and v be two vertices in a directed graph G. Using method (2) of handling dangling nodes of graphs (described in Section 2.4.1), the transition matrix S is constructed according to the following rules:

- If $\text{od}_G(u) > 0$, i.e., the vertex u is not a dangling vertex, all outgoing edges are weighted evenly as follows:

$$s_{uv} = \begin{cases} \frac{1}{\text{od}_G(u)} & \text{if } (u,v) \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

- Otherwise, let $s_{uv} = \frac{1}{n}$ for all $u \in V(G)$ where $\text{od}_G(u) = 0$. This distributes an even weight to each edge originating at the dangling vertex u and terminating at each node in the graph (including the dangling vertex itself).

This approach ensures that the transition matrix S is stochastic, since $0 \leq s_{ij} \leq 1$ and $S\mathbf{1} = \mathbf{1}$, and therefore satisfies Equation 2.4.2.

PageRank values cannot be computed from the matrix S since solving the equation $\boldsymbol{w}^T S = \boldsymbol{w}^T$ can result in multiple eigenvectors $\boldsymbol{w}$ associated with eigenvalues of magnitude 1, where each element of $\boldsymbol{w} \geq 0$ and $\boldsymbol{w}^T \cdot \mathbf{1} = 1$ [25].

The PageRank algorithm applies a simple solution by using a convex combination of S and an initialization vector $\boldsymbol{r}$, where each element of $\boldsymbol{r} > 0$ and $\boldsymbol{r}^T \cdot \mathbf{1} = 1$. The vector $\boldsymbol{r}$ typically contains the value of $\frac{1}{n}$ for each element where n is the number of vertices in the associated graph. The resulting matrix is defined as follows:

$$P = (1 - \alpha)\mathbf{1}\boldsymbol{r}^T + \alpha S \tag{2.5.1}$$

where α is the damping factor and is further discussed below. Using a convex combination of $\mathbf{1}\boldsymbol{r}^T$ and S ensures that the matrix P is irreducible since now all nodes are directly connected to each other, keeping the transition matrix stochastic and making it irreducible, by definition, and aperiodic (see Definition 2) since a period of $k = 1$ exists for each node.

Since P fulfills the ergodicity requirement a unique eigenvector exists with a magnitude of 1 [25]. This unique left eigenvector $\boldsymbol{x}$ from $\boldsymbol{x}^T P = \boldsymbol{x}^T$ can be computed using the power method which converges to $\boldsymbol{x}$ (see Section 2.5.2).

2.5.1 The Damping Factor α

The damping factor of the PageRank algorithm has multiple uses and implications. Firstly, it is used to make the transition matrix of the Markov chain irreducible so that a unique stationary distribution can be computed. If the damping factor $\alpha \in [0, 1)$ then the transition matrix is irreducible. The closer α is to 0, the more random restarts occur. In contrast, when $\alpha \to 1$ more focus is placed on the underlying network structure and the more accurately the underlying graph is modelled. Using the analogy of a random researcher, the smaller the damping factor, the more likely the random researcher stops following citations and chooses a new random paper. Conversely, if $\alpha = 1$, then the random researcher does not stop a search until reaching a dangling vertex.

Secondly, the damping factor has an impact on the ranking results as well as on the convergence speed of the computation, which in turn impacts the computation times of PageRank. According to Langville and Meyer [26], $log_\alpha \delta$, where δ controls the precision of the computation, can be used to roughly predict the number of iterations required for PageRank to converge. Therefore, as $\alpha \to 1$, more iterations are required and in addition increases numeric instability which means that the results of the computation do not accurately reflect the characteristics of the underlying graph.

Moreover, the nature and structure of the hyperlink graph of the Internet (webgraph) and academic citation networks differ in important ways. Webgraphs are dynamic since hyperlinks can be added or removed by updating webpages at any point in time. Outgoing edges of vertices in a citation network are fixed since references cannot be added to a paper after it has been published. In addition, webpages can be deleted from the webgraph but papers, once integrated into the academic corpus, are permanent. Vertices in a citation network can only acquire new incoming edges over time by citations from papers that are published at a later point in time.

This introduces an inherent time variable in citation networks which has to be considered separately and influences the use of the damping factor. More precisely, α controls the distribution of the ranking scores over the publication years of papers in citation networks. The smaller the value of α, the more evenly the scores are distributed over the years. Alternatively, a larger value of α has the effect that older papers are prioritised and receive larger ranking scores on average compared to recently published papers. The effects of varying damping values of PageRank when applied to citation networks are discussed in more detail in Section 7.1.

Therefore, the sensitivity and the accuracy of modelling the underlying graph, as well as the score distribution over the publication years, have to be balanced and optimised for each citation network while taking computation times into account.

2.5.2 The Power Method

The power method, or power iteration, is an algorithm that computes an eigenvector $\boldsymbol{x}$ of a matrix P associated with its largest absolute (dominant) eigenvalue λ [25, p. 5]. In other words, the power method solves the equation $P\boldsymbol{x} = \lambda\boldsymbol{x}$.

The algorithm starts with an initial vector $\boldsymbol{x}_0$, which can be a random vector or an approximation of the dominant eigenvector. The computation is then described by the following iteration:

$$\boldsymbol{x}_{k+1} = \frac{P\boldsymbol{x}_k}{\|P\boldsymbol{x}_k\|} \tag{2.5.2}$$

The sequence $\boldsymbol{x}_k$ only converges to the eigenvector associated with the dominant eigenvalue of P if the following two conditions hold:

- The matrix P needs to have one eigenvalue that is strictly larger than all its other eigenvalues.
- The initial vector $\boldsymbol{x}_0$ must contain a non-zero component that points in the direction of the eigenvector associated with the dominant eigenvalue.

Luckily, the eigenvector associated with the dominant eigenvalue coincides with the steady-state distribution of a Markov chain, as long as the transition matrices are constructed from citation graphs as described in Section 2.4.1. The power method is an efficient algorithm for computing this eigenvector, given that the transition matrix is very sparse.

Citation networks are inherently very sparse and even when the rows of dangling nodes are replaced with filled row vectors to make the transition matrices ergodic, the power method remains very efficient. This is shown in the following paragraphs.

Let S be the transition matrix that is used to model the random walks of researchers. The matrix S is constructed from two components. Firstly, the adjacency matrix A that models the connectivity of the underlying graph structure. And secondly, the additional component that is required due to the fitting of the dangling nodes. Therefore, the matrix S can be deconstructed as follows:

$$S = A + \boldsymbol{d} \cdot \boldsymbol{s}^T \tag{2.5.3}$$

where $\boldsymbol{d}$ is a vector containing ones for positions corresponding to dangling nodes and zeros otherwise. The vector $\boldsymbol{s}$ is a vector containing the weights for the added edges from the dangling nodes to all other nodes in the graph. For example, in the case of the weight being evenly distributed between the edges, $\boldsymbol{s}$ is filled with values equal to $1/n$, where n is the number of nodes in the graph.

Let P be the convex combination of the matrix S and the matrix $\mathbf{1} \cdot \boldsymbol{r}^T$ where $\boldsymbol{r}$ is the vector containing the probabilities of random researchers restarting their searches on a vertex. Therefore,

$$\begin{aligned} \boldsymbol{x}^T P &= \boldsymbol{x}^T \left[\alpha\left(A + \boldsymbol{d} \cdot \boldsymbol{s}^T\right) + (1-\alpha)\mathbf{1} \cdot \boldsymbol{r}^T\right] \\ &= \alpha\boldsymbol{x}^T A + \underbrace{\alpha\boldsymbol{x}^T \cdot \boldsymbol{d}}_{scalar} \cdot \boldsymbol{s}^T + (1-\alpha)\boldsymbol{r}^T \end{aligned} \tag{2.5.4}$$

From the above equation it is clear that the only computation that is not linear is the multiplication of the vector $\boldsymbol{x}^T$ by the matrix A. Fortunately, A is very sparse since $\overline{\mathrm{d}(v)} \ll n$, so that this computation is also $O(n)$.

2.6 Chapter Summary

This chapter outlined and discussed the various academic fields that are relevant to the topics presented in this thesis. The above chapter thus established the relevant academic context in which this research is situated.

In addition, domain-specific terms, the mathematical notation used throughout this document, and background information on Markov chains and the PageRank algorithm were presented and discussed in this chapter.

Chapter 3

Literature Review

In this chapter a review on the history of citation analysis and the current research on this topic is presented in order to provide the reader with the relevant background information. In the early stages of citation analysis, only citation counts were used as a proxy to measure the academic quality of articles, authors and journals. Simple metrics using citation counts, which are discussed in more detail in Section 3.1, were used to rank these entities accordingly.

This spurred a lot of debate within the academic community and since then citation analysis has been surrounded by a number of different viewpoints and opinions on how well citations can measure academic quality. In Section 3.2 these different viewpoints are presented and properties of papers that can or cannot be identified by citations are discussed. For example, self-citations of authors is a common practice in academia and can easily be identified. However, the question of what self-citations indicate remains open. In contrast, obliteration can occur to papers since their work has been firmly integrated into the general body of knowledge which results in less citations. This loss of citations, due to obliteration, is one example of citation behaviour that citation counts cannot identify and account for.

Since the emergence of digital academic libraries and computers capable of indexing large amounts of citation information, the focus of citation analysis has shifted towards algorithmic approaches for calculating academic quality and impact. Therefore, a review of the current research that is related to ranking academic papers, authors and journals algorithmically is given in Sections 3.3 and 3.4.

3.1 The History of Scientometrics and Bibliometrics

Before the age of specialization in the academic community, there was no need for indexing the current knowledge corpus in the sciences [1]. As Gross and Gross point out, libraries contained the general information for scholars to receive a standard education. This changed in the beginning of the 1920s when universities started shifting their focus from undergraduate work toward graduate studies by offering advanced speciality courses due to the demand for a highly skilled workforce.

The authors also note that, as a result of this shift in the structure of tertiary education the need for librarians to identify the most important journals that cover most specialities became apparent. This became especially important for smaller universities because of their limited financial resources to sustain large collections of periodicals. The need to

rank and identify the appropriate journals became a crucial requisite for universities to successfully prepare students for graduate studies in speciality fields.

Gross and Gross [1] noticed this need and published an article in 1927 suggesting a simple ranking metric for journals by selecting a single representative base journal in a field and counting all references contained in articles in all issues of the journal's latest volume. The journals which were cited the most were then considered the most important for libraries to acquire since they were assumed to be representative of the current research field.

This approach was used for a long time and was never scientifically questioned until Brodman, in 1944, proved that the method used by Gross and Gross is based on false assumptions and that their results do not correlate with accrued results of expert opinions [27].

The assumptions of the method used by Gross and Gross are:

1. The value of a journal to a researcher is directly proportional to the number of times its articles are cited in the academic literature.

2. The journals used as the base for the computation are representative of the entire field.

3. If more than one journal is used as a base, all of them can be weighted equally.

Brodman did not supply a more adequate method for journal choosing and merely pointed out drawbacks of the Gross and Gross method. This shows how difficult and controversial it is to measure academic importance based on citation counts alone.

Nonetheless, the use of citation counts, the basis of most bibliometric analyses, remains a topic of debate. Furthermore, results based on citation counts have to be interpreted carefully. This is discussed in further detail in the following section in which possible measures that citations can provide and different interpretations of what citations convey are outlined.

3.2 What Citation Counts Can and Cannot Measure

There exists an ongoing debate in the academic community of how the impact of papers, the prestige of journals and conferences, and the prominence of university departments should be measured. The controversial question is: What exactly do citations of academic papers measure? Without any additional evidence, what is the value of a citation? In his 1979 article "Is citation analysis a legitimate evaluation tool?", Garfield [8] tries to summarise what citation counts can and cannot measure and collects the different viewpoints in the scientific literature on the various aspects of academic citations. In this section the most important opinions are reiterated with references to newer literature in order to find answers to the following question:

> What is the relationship between a paper's citation count and its quality?

It is also important to define exactly what quality means in the context of academic articles. For example, a high quality paper does not necessarily indicate high impact. On the other hand, a high-impact paper does not presuppose quality. Therefore, the difference between impact and quality of papers and its relationship to citation counts are also discussed in this section.

3.2.1 Do High Citation Counts Indicate Quality Work?

The first debated question is whether a high citation count of a paper equates to quality work and high-impact research. Some believe that this is not true because a paper of low quality or one that contains incorrect results can also achieve a high citation count because it draws a lot of criticism.

Others argue that this situation is unlikely because, in general, academics tend to be reluctant to go to all the trouble to refute inferior work. It is more likely that bad material is bypassed and simply "dies", never to be cited again. A formal rebuttal which leads to increased citation counts only becomes necessary if incorrect results stand in the way of further development of a subject or if they contradict work in which someone else has a vested interest. Some even go further and state that if effort is invested into criticizing work, the work must be of some substance. Similarly, some researchers are of the opinion that formal refutations are also constructive and can clarify, focus and stimulate the research surrounding a certain subject. They argue that high citation counts are not a measurement of how many times an individual was right but rather that it measures the level of contribution of an individual to the practice of science.

Martin [16] argues that multiple indicators should be used to evaluate research and differentiates between research *quality*, *importance* and *impact*. He defines quality as

> a property of the publication and the research described in it. It describes how well the research has been done, whether it is free from obvious 'error', how aesthetically pleasing the mathematical formulations are, how original the conclusions are, and so on.

It is important to note that quality of academic publications is a relative measurement, requiring the judgement of other persons and is therefore dependent on personal attributes such as the cognition, opinion and social background of reviewers.

Martin defines the importance of a publication as

> its *potential* influence on surrounding research activities – that is, the influence on the advance of scientific knowledge ....

In contrast, he defines the impact of a publication as

> its *actual* influence surrounding research activities at a given time. While this will depend partly on its importance, it may also be affected by such factors as the location of the author, and the prestige, language and availability of the publishing journal.

He argues that citation counts are an indicator that best assesses a publication's impact rather than its quality or importance but that citation counts are only a partial indicator of impact and that other factors such as communication practices, author visibility and employing organisation have to be assumed significant [16, p. 7].

3.2.2 The Impact of Self-Citations

The term self-citation usually refers to a citation where at least one distinct author co-authored both the citing and referenced articles. Self-citation also occurs for research groups, journals and universities; in this section, the term *author self-citation* is used for a citation where the citing and cited papers have at least one author in common.

Methodologically, there are two types of self-citation rates; synchronous and diachronous [28]. On the one hand, synchronous author self-citations are references from within an article to another paper written by the same author. In order to obtain an author's synchronous self-citation rate, only information about the author's published work is required since it is the percentage of self-citations within the reference lists of the author's articles. Diachronous author self-citations, on the other hand, are in the set of citations that an article receives. In other words, a list of all papers that refer to the author's work is needed to compute the author's diachronous self-citation rate. Therefore, a citation index is typically required to find the referencing papers and to compute the diachronous self-citation rate of authors.

On the topic of author self-citation, opposing opinions also exist within the academic community [8, p. 4]. Some believe that self-citation manipulates citation rates. Others believe that self-citation and even team self-citation is very reasonable because it is more of an indication of a narrow speciality where scientists tend to build on their own work and that of collaborators.

Phelan [29, p. 8] argues, for example, that self-citation is an acceptable practice since it conveys the incremental nature of an individual's research and that it bears valuable information. Nonetheless, Phelan concludes that author self-citations should be excluded when performing citation analysis at author level but that they do not have a large impact on citation analysis at aggregated levels, such as at university level or country level.

On the basis of this, Aksnes [30] analyses self-citation rates in the Norwegian scientific literature between the years 1981 and 1996 using a sample of over 46 000 publications. He finds that 21% of all citations are author self-citations and that there exists a strong correlation between the number of authors of a paper and its self-citation rate. Furthermore, he finds that self-citations only contribute to a minor increase in the overall citation counts of multi-authored papers. He also identifies that self-citation rates vary significantly between academic disciplines. For example, the self-citation rate in clinical medicine is only 17% while the fields with the highest percentage of author self-citations are chemistry and astrophysics with 31% each.

Lastly, Aksnes concludes that if citation counts are used as research impact indicators, self-citations have a larger influence on the results when the time period of observation after publication used is short [30, p. 8]. For example, if citations are only counted for two years after the initial publication of papers, self-citations have a significant impact on citation rates which decreases the longer time period of observation is used.

3.2.3 Varying Citation Potentials

Another topic with differing views is the varying citation potentials in different academic fields. Citation potential is the likelihood of a paper receiving a citation at a certain point in time. A lot of different aspects may contribute to the probability of a paper getting cited. For example, the research field that the paper deals with, the venue at which the paper is published, or the quality of the paper may influence the likelihood of citation.

According to Garfield [8], some researchers are of the opinion that methodological advances are less important than theoretical ones. These researchers believe that citation counts cannot be a valid measure because they favour those who develop research methods over those who theorize about research findings. In general, method papers are not highly cited but this is also field dependent. Academic fields that are more oriented to methodology tend to be cited more. Instead of the importance or impact, the quality that citation counts measure is actually the utility or the usefulness of a paper to a large

number of people or experiments. On the other hand, the citation count of a work does not necessarily say anything about its elegance or its relative importance to the advancement of science or society. It only says that there are more people working on a specific topic than on another topic and therefore citation counts actually measure the activity of a topic at a certain point in time.

Alternatively, the number of publications of authors could be used to measure their contribution to scientific knowledge. This is also difficult since most publications only add small incremental additions to knowledge, while only a few make major contributions [16, p. 5]. The problem is that neither citation counts nor publication counts alone can be used to measure the quality and impact of an author's work.

3.2.4 Where Citation Counts Fall Short

All the above-mentioned work in this section refers to aspects of papers that can be identified and measured by using only citation counts. For example, the impact of self-citations is measurable by analysing citations and the varying citation potential of different academic fields can be computed by using citation counts if the required meta-data is available. The output or value of these measurements simply depend on the context in which the citations were counted and the interpretation of what citations actually mean.

Other aspects are not reflected by pure citation counts and additional information is required to rank academic articles with methods that take these aspects into consideration. These points are very important since different techniques of calculating a paper's importance have to be devised that are not only based on pure citation counts in order to assist or replace expert opinions.

Firstly, work that is very significant but too far ahead of the field to be picked up by others will go unnoticed until the field catches up. Citation counts will not identify significance that is unrecognized by the scientific community. They only reflect the community's work and interest [8]. As mentioned before, Martin [16] distinguishes between research quality, impact and importance. When citation counts are used as a measurement of impact and interpreted as such instead of a quality measure, then the criticism surrounding work that goes unnoticed but is of high quality can be avoided.

Secondly, obliteration is another issue that is not measurable by merely looking at a paper's citation counts. Obliteration occurs when some work becomes so generic to a certain field or has become so integrated into the body of knowledge that researchers neglect to acknowledge it with a citation. It is obvious that obliteration occurs to every work that is of high quality or that had a great impact in a certain field [8]. The problem is that obliteration can either occur shortly after publication or slowly over time which in turn will result in a high citation count and will render additional citations redundant. Either way, obliteration is not reflected in the citation counts of papers.

Another aspect of papers where additional information is required are the impact factors of the publication venues of citing or cited papers. Here it is very difficult to decide how individual citations should be weighted if information about publication venues is known. Should a citation to a paper published in a renowned journal, such as Nature, count more because it indicates excellent work? On the other hand, should the citation not count less because of the high visibility of the renowned venue? What is even more important is the question of whether the impact factor of the venue of the citing paper is as important as the impact factor of the venue of the referenced paper. For example, a reference from an article that is published in the journal Nature clearly indicates that the cited paper is of high quality.

Martin [16], for example, argues that a paper of high quality in a small and unpopular field or published at a small journal may have relatively low impact. On the other hand, an article published by a renowned author may have more visibility and therefore have higher impact with more citations, regardless of the paper's quality.

One last aspect closely related to pure citation counts which should be mentioned is that journal cross-citation is also important. Different academic fields have varying citation potentials which are dependent on aspects such as how quickly a paper will be cited, how long the citation rate will take to peak, the average length of reference lists in a certain field and how long a paper will continue to be cited. Figure 5.8 in Section 5.4 shows the varying citation rates of papers since their publication for different academic domains.

3.2.5 The Impact of Article Visibility on Citation Counts

Open Access (OA) is a term which is not well defined [31] but generally describes the principle of articles being visible online and easily accessible. More specifically, OA articles are digital, online, free of charge and free of most copyright and licensing restrictions [32]. According to Suber [31], OA can be classified into "gratis OA" which removes price barriers and "libre OA" which removes price barriers and at least some permission barriers.

With the emergence of online libraries and ease of access for scholars to obtain OA articles, certain new citation behaviours have been identified that influence which papers are more likely to get cited. For example, Lawrence [33] shows, using computer science articles from conference proceedings, that articles published online and free of charge are cited significantly more often than articles that are secured behind a paywall or are not made available online. Similarly, Brody and Harnad [34] show that physics articles that are submitted to pre-print and later published in peer-reviewed journals receive an up to 400% higher citation count than articles that were not published on ArXiv, a repository of OA digital pre-prints of scientific papers [35].

By analysing the access logs of the NASA Astrophysics Data System Digital Library, Kurtz [36] shows that articles published in journals that have restrictive access policies have half the chance of being read by researchers compared to articles in journals with more liberal access policies.

In a later study, Kurtz *et al.* [37] propose three potential aspects of journal article publishing policies that could explain the impact of increased citation counts, as identified by Lawrence [33] and Brody and Harnad [34], and try to verify them based on data from the field of astronomy.

The first aspect pertains to the relationship between increased citation counts and OA articles. Kurtz *et al.* [37] find no evidence to support the hypothesis that articles that are not restricted by a paywall system are cited more frequently. They argue that an astronomer who publishes articles has to have obtained a certain authoritative position and therefore has no restrictions to read the journals. It should be noted that no evidence is given to support this argument. Furthermore, their findings are based on publication data restricted to the field of astronomy and cannot be generalized to all academic fields and journals because of two main reasons. Firstly, journals in different academic fields may have different preferred access policies [38, p. 4] and secondly, the total cost of subscribing to the main journals in different fields can vary because of the fields' sizes.

The second aspect that Kurtz and his colleagues investigate is the early access attribute of articles that are published as pre-prints openly before appearing in journals. For the field of astronomy, they find that the correlation between open articles, published at

ArXiv, and a higher citation count cannot be attributed to this early access attribute alone even though the open articles have more than twice the probability of getting cited [37].

They conclude that the correlation between OA articles and a higher citation count is caused by a combination of the early access attribute and a selection bias of the authors. They show that researchers can boost citation counts of their articles by self-promoting favourite articles by means of posting them on personal websites or public forums.

Moed [39] agrees with the statement that the two factors that account for the increased citation count of OA articles are firstly the preview effect and secondly the free access to online self-published articles by the authors themselves.

Davis [40] conducted a randomized controlled trial on OA articles versus subscription-based articles in 36 journals in different academic domains. He found that articles that are openly accessible do find a wider audience with more resource downloads but that it does not have a significant impact on articles' citation counts and also does not impact the timeline of accruing citations.

It should be noted that the above-mentioned studies conducted on the bias of OA articles are based on citation indices that have a selection bias since the sources are curated and only include international and high-impact journals in their fields. Authors of articles published in these journals will typically have access to the journals anyway, which impacts the studies of open access. The open access impact on citation counts cannot be measured by using these types of databases. It would be interesting to conduct the same studies on data sets that include national or less renowned journals. However, this is beyond the scope of this thesis.

3.2.6 Citation Analysis, Data Quality and Coverage

Citation analysis is very dependent on the coverage and the quality of data sources since it is based not only on citations but also on the type of papers that are indexed. For example, some data sources include editorials, reviews and technical reports, while others do not. Moreover, the update frequencies of the databases and the included languages of papers also vary [41]. Using the same citation analysis methods on different data sources and comparing the results is tricky because of discrepancies in coverage and because the paper type is not always specified.

Zhang [42], for example, uses a sample of 25 randomly selected computer scientists from Canadian universities and shows that Scopus[1] identifies 90% of their publications while Web Of Science[2] only identifies 55%. Citation counts also differ substantially, where Scopus retrieves 65% more citations. This is understandable due to the higher number of citable items in the Scopus database but would skew results dramatically if citation counts are directly compared between data sources. In addition, Zhang finds that Web Of Science contains a higher percentage of journal articles than conference proceedings compared to Scopus.

Similarly, Franceschet [45] compares the number of publications and the citation counts of authors who belong to a computer science department of an Italian university. He finds that Google Scholar has five times the publication counts and eight times the citation

[1]Scopus is a multi-disciplinary bibliographic database containing abstracts and citation information of peer-reviewed journals and conference proceedings [43].

[2]Web Of Science is a scientific citation index of multi-disciplinary journals, books and conference proceedings [44].

counts compared to Web Of Science. However, Franceschet also shows that rankings based on citations do not change significantly when these two data sources are used.

Kulkarni *et al.* [46] analyse the citation characteristics of 328 medical papers published in three medical journals and compare their characteristics based on citation data from Google Scholar, Scopus and Web Of Science. They find that Google Scholar and Scopus find more citations and that Scopus finds more citations from non-English papers compared to Web Of Science. In addition, Google Scholar has significantly less citations to group-authored articles compared to the other two data sources.

Chapter 5 briefly discusses the quality and the properties of the data sets used for the experiments in this thesis.

3.3 Ranking Publications

In recent years, and due to automated citation indexing, bibliometric research has shifted towards the citation analysis of large scale citation networks and has allowed researchers to apply advanced methods for pattern recognition, knowledge discovery and impact measurements. With the launch of online citation indexing services such as CiteSeer [47], Google Scholar [48] and Microsoft Academic Search [5], and, in general, the access to large publication data sets, more advanced models of citation analysis have been proposed. In this section, some of these methods are described with a focus on algorithms that use citation networks as basis for their computations and calculate impact scores for individual papers.

The PageRank algorithm was first devised by Brin and Page [4] in 1998 to rank websites according to their importance by calculating an impact score based on the number of referring hyperlinks. The more hyperlinks from other important websites point to a particular website, the higher the score of the website.

This idea of the PageRank algorithm has been applied to academic citation networks frequently. For example, Chen *et al.* [12] apply the algorithm to all American Physical Society publications between 1893 to 2003. Their research shows that there exists a close correlation between a paper's number of citations and its PageRank score but that important papers, based purely on the authors' opinions, are found by the PageRank algorithm that would not have easily been identified by looking at citation counts only. Chen *et al.* use the basic PageRank algorithm as given by Equation 4.2.3 with a damping factor $\delta = 0.5$ instead of 0.85. They argue that entries in the bibliographies of papers are compiled by authors by searching citation paths of length two on average. Choosing a damping factor of 0.5 leads to an average citation path length of 2 in the PageRank model which seems more appropriate for citation networks. They base this choice on the observation that about 42% of the papers that are referenced by a paper A have at least one reference directly to another paper that is also in the reference list of A. This value was computed from a data set containing physics publications and may be different for other academic domains.

Using the same data set and in addition a citation data set of all journals published by the American Physical Society, the authors of [13] devise an algorithm, called CiteRank, that simulates the flow of traffic through citation networks from recently published papers to older papers following citations. The CiteRank algorithm takes the publication dates of papers into consideration to account for the aging characteristics in citation networks. The results of the CiteRank algorithm are compared to the unmodified PageRank algorithm by looking at outliers and discussing the reasons for either a high CiteRank or PageRank

score. Similarly to the results shown by Chen *et al.* [12], the discussion on the effectiveness of their proposed algorithm is subjective to the authors opinions. The details of the CiteRank algorithm are given in Section 4.2.3.

Similarly, Hwang *et al.* [14] modify the PageRank algorithm by incorporating two additional factors when calculating a paper's score. Firstly, the age of a paper is taken into consideration and secondly, the impact factor of the publication venue of a paper is also included in the computation. The algorithm was proposed in an article called "Yet Another Paper Ranking Algorithm Advocating Recent Publications". For brevity this algorithm is referred to as YetRank and is described in Section 4.2.5.

Dunaiski and Visser [15] propose an algorithm, NewRank, that also incorporates the publication dates of papers similar to YetRank. They compare the NewRank algorithm to PageRank, CiteRank and YetRank and find that it focuses more on recently published papers. In addition, they evaluate the algorithms using papers that won the "Most Influential Paper" award at ICSE (International Conference on Software Engineering) conferences and find that PageRank identifies the most influential papers the best.

Sidiropoulos and Manolopoulos [11] propose an algorithm that is loosely based on PageRank. The authors call their algorithm SceasRank (Scientific Collection Evaluator with Advanced Scoring). SceasRank places greater emphasis on citations than the underlying network structure compared to PageRank. Two additional variables are introduced that control the impact of indirect citations and the weight that should be associated with citations that originate from papers that have no citations themselves.

Sidiropoulos and Manolopoulos use a data set of Computer Science papers from the DBLP library [7] and compare different versions of the SceasRank algorithms with PageRank and pure citation counts. They evaluate the algorithms using papers that won impact awards at one of two venues. Firstly, papers that won the 10 Year Award [49] at VLDB (Very Large Data Base) conferences, and secondly, the papers that won SIGMOD's (Special Interest Group on Management of Data) Test of Time Award [50] are used as evaluation data to judge the ranking methods in ranking important papers. Their results show that SceasRank and PageRank perform the best in identifying important papers but that using citation counts is very close to those methods.

They also rank authors by using the best 25 papers of each author and use the "SIGMOD Edgar F. Codd Innovations Award" [50] as evaluation data. Their results show that SceasRank performs equally well compared to PageRank and improves over the method of simply counting citations to find important authors.

3.4 Ranking Authors and Venues

The idea of an impact factor for journals was first introduced by Garfield in 1955 [2, p. 4] by indexing bibliographies automatically and using this information to rank journals. The Journal Impact Factor was then formalized to measure the average citation frequency of articles published in a journal in a certain period of time [51]. It was devised to overcome the problem that smaller yet important review journals of a speciality subject matter might not be selected if a ranking scheme is solely based on the total number of publications or total citation counts. It computes a relative importance number that can be used to compare journals, and consequently conferences, within the same academic field [2]. The official Journal Impact Factor is computed by Thomson Reuters, formerly known as the Institute for Scientific Information (ISI).

According to Garfield [52] the Journal Impact Factor reduces the bias of total citation counts of larger journals over smaller journals or journals that publish less frequently. In addition, it does not prefer newer journals over older journals. He concludes that the larger the number of articles published in a journal, the more citation counts the journal will accumulate.

Currently the h-index method, developed by Hirsch [3], is the de facto technique of calculating quality and impact of a researcher's work in the academic community. The h-index is defined as:

> A scientist has index h if h of his/her N_p papers have at least h citations each, and the other $(N_p - h)$ papers have no more than h citations each.

This metric is only applicable to compute scores for an author or a group of authors and not for individual papers. Therefore, the h-index can only be used to compute scores for journals, conferences, individual authors or academic departments. For a more detailed discussion on the h-index see Section 4.1.3. This h-index is a very simple metric that is based on the citation counts of papers directly. The h-index value is dependent on an author's most cited papers and the number of citations that they have received in other publications. Therefore, the h-index tries to measure both the quality (number of citations of most cited papers) and quantity (the number of papers published over the years) of an author's work. As with all other citation analysis methods that use citation counts directly, the h-index does not account for a lot of the characteristics described above and features that are common to citation networks. For example, the h-index does not consider the number of authors of a paper, the varying citation potentials of different academic fields and is dependent on the total number of publications of authors.

Bollen *et al.* [53] use a Weighted PageRank algorithm on a journal graph to compute journal scores based on the idea that the output of the PageRank algorithm focuses more on the prestige of journals compared to the output of the Impact Factor which computes rankings that reflect more the popularity of journals. They find that, in general, the Impact Factor ranks review journals favourably compared to the PageRank algorithm.

Since the Impact Factor is known to have biases when it is used to compare journals across different academic disciplines [54; 55], they compare the output of the Impact Factor and the Weighted PageRank algorithm for the domains of computer science, physics and medicine individually. They conclude that for physics the Weighted PageRank prefers journals that are generally favoured by domain experts [53, p. 10] and that for computer science it prefers journals that are heavily subject-focused. For medicine journals, they find that the notion of prestige and popularity is more intertwined than in computer science and physics.

In addition, Bollen *et al.* [53] propose a metric called *Y-factor* which is a combination of the Weighted PageRank and the Impact Factor results by multiplying the two values for each journal. They draw no conclusions about the results of this metric except that it is comparable to the h-index when applied to medicine journals [53, p. 7].

The Eigenfactor project, created by Bergstrom *et al.* [9], ranks academic journals and has recently gained a lot of attention. The journal scores are computed using a PageRank-like algorithm on a journal citation graph and have been included in the Thomson Reuters "Journal Citation Report" [56] since 2007.

The Eigenfactor Metric computes two scores for journals. The first is the Eigenfactor score, indicating the total importance of a journal which is the sum of all scores of articles published within that journal. Therefore, larger journals that, on average, publish

more articles a year will have greater Eigenfactor scores [57]. The second score that the Eigenfactor Metric calculates is the Article Influence score. This score is intended to measure the influence of a journal by averaging a journal's score by the number of articles it publishes. Therefore, the Article Influence scores of journals can be compared to their Impact Factor scores.

3.5 Chapter Summary

In this chapter various approaches of ranking journals, authors and papers are presented as found in the literature, from early journal ranking to current algorithmic approaches to rank journals and papers.

In addition, the feasibility and impracticability of using citation counts to measure quality, importance or impact of papers are put forward and discussed. The bottom line is that, without additional information, the quality of papers is very difficult to compute and that rankings are more likely to convey the impact or visibility of papers than their intrinsic academic quality.

It follows that, when evaluating individual papers, citation counts can only be used as an aid to provide an objective measure of the utility, impact or popularity of academic work. They say nothing directly about the quality of the work and nothing about the reason for the utility or impact of the work.

It should also be noted that citation analysis results can only be as good as the data on which it is performed. Also, comparing results of citation analyses and impact metrics from different data sources is difficult and the coverage, accuracy and content of the data sources have to be taken into consideration. The effect of discrepancies between data sources are normalised to a certain extent if citations are used for rankings. Intuitively this can be assigned to the fact that the lack of coverage, for example, impacts every researcher to roughly the same extent. However, if a researcher predominantly publishes at conferences that are not indexed by a certain citation index then it has a bigger impact on his or her rankings.

Lastly, it should be mentioned that alternative approaches for measuring a researcher's impact in the scientific community have been studied that do not rely on citations. Other usage data that can now be indexed, such as the number of downloads of an electronic article or the number of page views of a publication, can be used as indicators for importance or impact. However, any discussion of these alternative approaches would go beyond the scope of this thesis.

Chapter 4

Ranking Methods

This section describes various citation analysis methods and ranking algorithms that are closely related to, or directly used in, the research presented in this thesis. The chapter is organised into three sections, in each of which a different group of ranking methods is discussed.

The first section discusses well known and often used methods that compute scores for publication venues using citation analysis. Citation analysis, in the traditional sense of bibliometrics, is undertaken on data sets that contain information about articles and their references by counting citations and looking at citation distributions of venues. The methods introduced in the first section are mostly used either to rank academic journals and conferences or to compute impact scores for authors. The results of the methods have varying meanings and depend on the use of the methods and interpretation of the results. However, the methods described in this first section merely use pure citation counts as basis for their computations.

Citation information can be augmented to create citation networks which can include additional information such as author names, the publication dates of papers and the venues where papers were published. The algorithms discussed in the second section make use of this additional information. They are based on the PageRank algorithm or similar models of traffic and consider the structure of entire citation networks as the basis for computing scores for individual academic publications. The second section gives a brief overview of the current approaches that are proposed in recent literature for computing scores for individual academic articles.

Lastly, the third section of this chapter discusses other algorithms based on entire citation networks that are used to compute scores for publication venues. This section describes how algorithms described in Section 4.2 can be adapted to compute scores for publication entities such as journals or authors.

4.1 Counting Citations

4.1.1 The Journal Impact Factor

The definition of the Journal Impact Factor that is currently used by Thomson Reuters is the following [52]:

> In a given year, the Impact Factor of a journal is the average number of citations received per paper published in that journal during the two preceding years.

In order to generalise the formulation of the Journal Impact Factor, two time frames have to be defined. Firstly, the *census window* (CW) is a time frame that is defined to include all the papers whose outgoing citation should be considered. Secondly, the *target window* (TW) is a year range directly before the census window. All papers published in journals during the target window are potential citable items and references to these papers are used for measuring the importance of journals. In other words, all references originating from papers in the census window and citing papers in the target window are considered when computing impact factor scores for journals.

The census window and target window size, as defined by Thomson Reuters [52], are one and two years, respectively. For example, for the computation of the 2013 Impact Factor scores of journals, the year ranges [2011; 2012] and [2013; 2013] are used for the target window and the census window, respectively.

Let $\mathcal{P}(v, (t_1, t_2))$ be the set of papers that are published by venue v during the time frame $[t_1; t_2]$. Furthermore, let $G(V, E)$ be the underlying citation network with the associated set of venues $\mathcal{V}$. The following equation denotes the number of citations from any paper in $\mathcal{V}$ during the CW to papers that fall within the TW and are published at venue v:

$$\text{Cited}(v, CW, TW) = \sum_{\{(p_i,p_j)\in E | p_i \in \mathcal{P}(\mathcal{V},CW) \wedge p_j \in \mathcal{P}(v,TW)\}} w(p_i, p_j) \tag{4.1.1}$$

If the Impact Factor for a journal were measured by using the above equation, then venues that publish a larger set of papers would be unfairly advantaged since they would have more citable items which is the set $\mathcal{P}(v, TW)$ in Equation 4.1.1. Therefore, the value is normalised by the number of articles associated with a venue during the target window as described by the following equation:

$$IF(v, CW, TW) = \frac{\text{Cited}(v, CW, TW)}{|\mathcal{P}(v, TW)|} \tag{4.1.2}$$

It should be noted that the Impact Factor is dependent on the citation rate of academic disciplines and therefore should not be used to compare venues that are from different domains. For example, assume that the sizes of two disciplines A and B are the same but that the average citation rate of A is much larger than B. Then $\mathcal{P}(v_A, TW) \approx \mathcal{P}(v_B, TW)$ but $\text{Cited}(v_A, CW, TW) \gg \text{Cited}(v_B, CW, TW)$ independent of the average impact of the disciplines.

4.1.2 The *i10*-index

The *i10*-index is a simple author impact measure developed by Google and introduced in 2011 on the Google Scholar website. An author has an *i10*-index value of i if the author has published i papers that have received at least 10 citations each [58]. Intrinsically, the *i10*-index only measures the impact of an author and is highly dependent on publication counts of authors.

4.1.3 The *h*-index

The *h*-index is a relatively new method developed by Hirsch [3] and was first published in 2005. It was developed for measuring the quality of theoretical physicists' research output but has since gained a lot of popularity in the academic community for computing the impact of researchers in general.

The h-index is based on citation counts solely and considers the distribution of citations of a researcher's publications. The h-index is defined as follows:

> An author has an index h if their h most-cited publications have h or more citations each.

More formally, let $\{p_1, p_2, p_3, \ldots \mid \mathrm{id}(p_i) \geq \mathrm{id}(p_{i+1})\}$ be an author's set of papers that is sorted in descending order of the number of citations. The h-index is then computed by stepping through this set and finding the largest value for h such that:

$$h \leq \mathrm{id}(p_h) \tag{4.1.3}$$

The h-index tries to improve on simply counting the total number of papers and the total number of citations that an author has received since the total number of papers does not measure the impact of the work and the total citation count of an author can easily be skewed by co-authoring a small number of highly cited papers which does not accurately reflect the authors overall contribution to science.

For example, assume that an author has published 10 articles, each of which has received only a single citation. The author's h-index is 1 indicating that the author's work is not of significant importance. Similarly, an author that has only published a single article that has received ten citation also only has an h-index of 1 showing that the contribution of the author to the academic corpus is small.

The main disadvantage of the h-index is that it is accumulative and does not decrease over time even if an author does not contribute to the research corpus anymore. Analogously, the h-index increases with the accumulation of citations. Therefore, it is dependent on the number of years since a researcher has published papers. Similarly, the h-index value is bounded from above by an author's publication count and therefore researchers with shorter academic careers are at a disadvantage.

In order to overcome the drawback of the accumulative property of the h-index, Google Scholar for example, lists two h-index values for authors. In addition to the standard h-index, an h-index value fitted to a time window of the last 5 years is given. Here, only citations that were received by all papers of an author in the previous 5 years are used to compute the h-index value. This alternative h-index value indicates whether an author has been actively contributing to the academic corpus in recent years.

It is very important to apply the h-index properly as proposed by Hirsch. Since there exist different citation conventions in various academic fields, researchers from different academic domains should not be compared using the h-index. Hirsch [3] identifies, for example, that high h-indices are much higher in social science than in physics.

Intrinsically, the h-index cannot be computed for a single publication since it is based on a set of papers associated with the entity for which the h-index value is computed.

In addition, the h-index value is highly dependent on the coverage and accuracy of the data set that is used. Franceschet [45] shows that computer scientists belonging to a university in Italy have, on average, a three times higher h-index when using Google Scholar citation data than using data from the Web Of Science. This is true for any impact measurement that is solely based on pure citation counts. But since the h-index is very dependent on the total number of an author's publications, the coverage of a data source is very important. Franceschet, for example, shows that rankings based on citations do not vary significantly but that rankings based on the h-index vary moderately.

Zhang [42] shows the same by using a sample of 25 randomly selected computer scientists from Canadian universities. Zhang shows that the average h-index of these authors is

2.1 times higher using the Scopus citation data compared to the Web Of Science database. However, the difference in the h-index is normalised to a certain degree when used for rankings. The two sets of rankings according to the h-index have a relatively high rank correlation (Spearman $\rho = 0.73$).

Meho and Rogers [59] conduct a similar study in which they compare the h-indices of 22 researchers in the field of human-computer interaction using Scopus, Web Of Science and Google Scholar. They find that Google Scholar, Scopus and Web Of Science compute an average h-index of 20.6, 12.3 and 8.0, respectively. However, Meho and Rogers also show that a high rank correlation (Spearman $\rho = 0.96$) is obtained when the Google Scholar citation information is compared to the combined data sources of Scopus and Web Of Science.

4.1.4 The g-index

The g-index was developed in 2006 by Egghe [60] and tries to overcome some of the drawbacks of the h-index. It is one of the more popular variations of the h-index.

> An author has a g-index value of g if their top g articles in sum have received at least g^2 citations.

Similarly to the h-index, let $\{p_1, p_2, p_3, \ldots | \mathrm{id}(p_i) \geq \mathrm{id}(p_{i+1})\}$ be an author's set of articles that is sorted in descending order of citation counts. The g-index is then computed by stepping through this set and finding the largest value for g such that:

$$g \leq \frac{1}{g} \cdot \sum_{i \leq g} \mathrm{id}(p_i), \tag{4.1.4}$$

Similarly to the h-index, the g-index measures two quantities. Firstly, it indicates the amount of research output an author has produced and secondly, it also gives an indication of the quality of the author's work. The g-index allows citations from highly cited papers to push up the g-index while not affecting the h-index therefore lowering the quality threshold. Therefore, g is at least the value of h but usually greater than the h-index value.

4.2 Paper Ranking Algorithms

In this section, ranking algorithms are described that compute relevancy scores of individual academic papers.

Let $G = (V, E)$ be a directed citation graph containing n papers and m references. When ranking papers by simply counting their citations, a ranking score $CCR(p)$ for each paper $p \in V$ can be calculated using the following equation

$$CCR(p) = \frac{\mathrm{id}_G(p)}{m} \tag{4.2.1}$$

resulting in scores between 0 and 1, with the norm of the result vector ($\|CCR\|_1$) equal to 1. For the remainder of this thesis the method of ranking papers according to their citation counts is referred to as CountRank (CCR). It should be noted that the citation counts of papers are normalised by the total number of citations in the network in order

for the CountRank scores to be comparable to the other ranking algorithms discussed in this section.

As mentioned in Section 3.2, a paper's citation count does not necessarily reflect its quality or importance to research. The drawbacks to using ranking techniques that merely count the number of citations of papers are summarised below:

P1: The first problem is that the publication years of papers are not considered. Recently published papers have not been around very long and therefore have not yet had a chance to accrue many citations. In contrast, papers that contain important work but were published a long time ago, might only be cited modestly because of a smaller scientific community [61].

P2: Another problem is that the age of citing papers is not taken into consideration. Citations from newer papers should count more than citations from older papers, especially if the aim is to identify currently important papers. For example, an old paper which is directly cited by a new paper indicates that it still bears current relevance.

P3: The third problem is that citations from highly cited papers should be regarded as more important than citations from less important papers.

P4: Citations from papers that were published at prestigious venues should carry more importance than citations from papers published at less renowned venues.

P5: Different academic fields have varying referencing conventions. These disproportionate citation potentials also depend on the size of the academic fields and the age of the disciplines.

The ranking algorithms described in this section, when applied to citation networks, try to address all or a subset of these problems which will be referred to by their names P1 through P5.

4.2.1 PageRank

The PageRank algorithm was developed to rank web pages according to their importance or relevance and uses the graph structure of the Internet as a basis for the computation [4]. The result of the PageRank computation is a probability distribution that represents the likelihood that a web surfer who is randomly clicking on links will arrive at a certain webpage. The probability that the random surfer stops following links and goes to a random page is given by the damping factor α.

Brin and Page [4] gave the following mathematical description of PageRank where the initial probability distribution at iteration $t = 0$ is given by

$$PR_0(p) = \frac{1}{n} \tag{4.2.2}$$

At each iteration of the algorithm the PageRank value for the webpage p_i is calculated using the following formula:

$$PR_t(p_i) = \frac{(1-\alpha)}{n} + \alpha \cdot \sum_{p_j \in N^-(p_i)} \frac{PR_{t-1}(p_j)}{\mathrm{od}(p_j)} \tag{4.2.3}$$

As mentioned previously in Section 2.5 the computation stops when the result vector converges to a predefined precision threshold δ:

$$\sum_{p \in V(G)} |PR_t(p) - PR_{t-1}(p)| < \delta \tag{4.2.4}$$

The analogy of a random surfer can be translated to fit the context of academic citation networks where, instead of a random surfer reading webpages and following hyperlinks to different webpages, a random researcher traverses a citation network by reading articles and following references to other articles by looking up references in bibliography sections.

All algorithms described in this section follow this analogy and are based on the same idea of calculating the predicted traffic to the articles in citation networks. The intuition behind these algorithms is that random researchers start a search at some vertices in the network and follow references until they eventually stop their search, controlled by a damping factor α, and restart their search on a new vertex. Since the result vectors of all ranking algorithms described in this section converge after a sufficient number of iterations, the computations stop when a predefined precision threshold δ is reached.

Therefore, the ranking algorithms differ in only two aspects:

- How are the random researchers positioned on the citation network when they start or restart their searches? Should a random researcher be randomly placed on any vertex in the network or does the random researcher choose a vertex corresponding to a recent paper with a higher probability?

- Which edge (citation) should the random researcher follow to the next vertex (paper)? Should the decision depend on the age of the citation? Should the impact factor of the venue at which the citing or cited paper was published contribute to the decision?

In the case of the standard PageRank algorithm the random researchers are uniformly distributed on the citation network, as given by Equation 4.2.2, and select the edge to follow at random (right hand side of Equation 4.2.3). In other words, all articles and references are treated equally and a random researcher does not have any preference in selecting a certain paper or following a reference to another paper.

The time complexity to compute one iteration of PageRank, where a PageRank value for each vertex is computed, is $O(n)$ as discussed in Section 2.5.2. Two values have to be stored in memory for each vertex in the network, the current PageRank value for each vertex and that of the previous iteration. Therefore, the space requirement for the PageRank algorithm is also $O(n)$.

The PageRank algorithm addresses P3, since it was developed to calculate the predicted traffic to a web page instead of simply counting the number of hyperlinks that point to a web page. Therefore, the PageRank algorithm seems like a good candidate to be applied on citation networks in order to rank papers. Additionally, it has been shown that the PageRank algorithm overcomes the problem of the varying citation potentials between different academic fields and negates the skewing effect that this problem has on the ranks of articles, therefore, addressing problem P5 [62].

The PageRank algorithm works well for the Internet's web graph but has certain drawbacks when used on citation networks. Unlike the web graph, citation networks are typically acyclic and have an intrinsic time arrow since papers can only cite older papers that have been published before. Furthermore, if researchers would randomly

follow citations without restarting their searches, given enough time, they would end up stuck at the old leaves of the citation network. Therefore, the aging effect [63; 64] of citation networks has to be considered. This aging effect can be counterbalanced by either modifying the PageRank algorithm and incorporating the publication dates of papers directly or, to a certain degree, by choosing an appropriate α value for the underlying graph data [12].

Additionally, the PageRank algorithm favours vertices that are contained within citation cycles. In bibliometric citation networks, citation cycles do not usually occur since papers can only reference papers that have been published already. Nonetheless, citation cycles can exist due to self-citations or erroneous data. See Figure 4.1 in Section 4.2.6 for an example graph that shows this behaviour.

The matrix notation of the PageRank algorithm is given in this paragraph for consistency and for easier comparison between the ranking methods based on traffic models.

Let A be the matrix of a graph G, where $a_{ij} = \frac{1}{\text{od}(p_i)}$ if $(p_i, p_j) \in E(G)$ and zero otherwise. Furthermore, let $\boldsymbol{d}$ be a vector with values $d_p = 1$ if the vertex corresponding to paper p is a dangling vertex and zero otherwise.

An iteration of the PageRank algorithm is then described by the following equation:

$$\boldsymbol{x}_t = \overbrace{\left[\underbrace{\frac{(1-\alpha)}{N} \cdot \mathbf{1} \cdot \mathbf{1}^T}_{\text{Random Restarts}} + \alpha \cdot (A^T + \underbrace{\frac{1}{N} \cdot \mathbf{1} \cdot \boldsymbol{d}^T}_{\text{Dangling Vertices}})\right]}^{\text{Stochastic Matrix } P} \cdot \boldsymbol{x}_{t-1} \tag{4.2.5a}$$

$$= \frac{(1-\alpha)}{N} \cdot \mathbf{1} + \alpha \cdot \left(A^T + \frac{1}{N} \cdot \mathbf{1} \cdot \boldsymbol{d}^T\right) \cdot \boldsymbol{x}_{t-1} \tag{4.2.5b}$$

where $N = n(G)$ is the size of the graph G. The above equation is one iteration of the approximation of the Power Method[1] to solve for the leading eigenvector of the stochastic matrix P. This definition of the PageRank algorithm uses solution 2 from Section 2.4.1 by adding N edges from each dangling vertex to all other vertices in the graph and evenly distributing the weight between the added edges. This is modelled by the "Dangling Vertices" term in Equation 4.2.5a, while the first part of the equation, $(1-\alpha)/n \cdot \mathbf{1}$, models the evenly distributed placement of random researchers when they restart a search.

The computation stops when the predefined precision threshold δ is reached, i.e.:

$$\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|_1 < \delta \tag{4.2.6}$$

4.2.2 SceasRank

The *Scientific Collection Evaluator with Advanced Scoring* (SCEAS) ranking method introduced by Sidiropoulos and Manolopoulos [11] and used in [65] is the PageRank algorithm as described above with alterations by introducing two parameters a and b. According to the authors, b is called the *direct citation enforcement factor* and a is a parameter controlling the speed at which an indirect citation enforcement converges to zero.

[1] See Section 2.5.2 for more information on the Power Method. When using the Power Method to approximate the leading eigenvector of the matrix P in Equation 4.2.5a, the term $P\mathbf{x}_k$ does not need to be normalised by the value $\|P\mathbf{x}_k\|$ because it is intrinsically equal to 1.

The following equation gives the definition of one iteration of the SceasRank algorithm

$$SR_t(p_i) = \frac{(1-\alpha)}{n} + \alpha \cdot \sum_{p_j \in N^-(p_i)} \frac{SR_{t-1}(p_j) + b}{\text{od}(p_j)} a^{-1} \tag{4.2.7}$$

Let A be the adjacency matrix of a graph G, where $a_{ij} = \frac{1}{\text{od}(p_i)}$ if $(p_i, p_j) \in E(G)$ and zero otherwise and let $\boldsymbol{x}_0$ be the initial probability distribution where $\boldsymbol{x}_0(p) = \frac{1}{n}$ for all $p \in V(G)$. Additionally, let K be a matrix that contains $k_{ij} = 1$ if $(p_i, p_j) \in E(G)$ and zero otherwise. Furthermore, let $\boldsymbol{d}$ be a vector with values $d_p = 1$ if the vertex corresponding to paper p is a dangling vertex and zero otherwise. The alternative notation for SceasRank is therefore:

$$\boldsymbol{x}_t = \frac{(1-\alpha)}{N} \cdot \mathbf{1} + \frac{\alpha}{a} \cdot \left(A^T + \frac{1}{N} \cdot \mathbf{1} \cdot \boldsymbol{d}^T \right) \cdot \left(\boldsymbol{x}_{t-1} + b \cdot K^T \cdot \mathbf{1} \right) \tag{4.2.8}$$

For $b = 0$ and $a = 1$ Equation 4.2.8 is equivalent to PageRank's formula given in 4.2.5a.

According to the authors, b is used because citations from papers with scores of zero should also contribute to the score of the cited paper. Furthermore, the indirect citation factor a is used to control the weight that a paper x citations away from the current paper has on the score and is a contribution that is proportional to a^{-x}.

In [11, p. 3] Sidiropoulos and Manolopoulos use SceasRank with two different sets of parameters and refer to them as SCEAS1 and SCEAS2. SCEAS1 assumes that $\alpha = 1$, $b = 1$, and $a = e$ while for SCEAS2 the parameters have the values $\alpha = 0.85$, $b = 0$, and $a = e$.

It should be noted that if a damping factor of $\alpha = 1$ is used, which is possible due to the parameter a, no N additional edges should be added to each dangling vertex in the graph since it skews the results. Instead, the algorithm reduces to the following and is referred to as SceasRank1 (SR1) in the following discussion.

$$SR1_t(p_i) = \sum_{p_j \in N^-(p_i)} \frac{SR1_{t-1}(p_j) + b}{\text{od}(p_j)} a^{-1} \tag{4.2.9}$$

Equation 4.2.9 does not model random researchers traversing a citation network that restart their searches and does not compute a steady-state distribution of a stochastic Markov chain. Rather it models random researchers that traverse the citation network until they stop their search due to the damping of the parameter a or because they reach the end of the citation network. Nonetheless, the vector $SR1$ still converges if $a > 1$ and $b \geq 0$, but not with a magnitude of 1 and depends on the values of a and b. Therefore, a normalisation step is required to ensure that the result vector has a magnitude of 1. The stopping criteria of $\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|_1 < \delta$ can then still be used.

Similarly to PageRank, the SceasRank algorithm addresses P3 and P5. In addition, SceasRank addresses P2 indirectly. P2 is the problem of taking the publication dates of papers into consideration. SceasRank addresses this problem to a certain degree by using the indirect citation factor a to controls the weight that citations carry along a citation chain. SceasRank's time and space complexity is $O(n)$ for each iteration of the algorithm. However, its main advantage is that it converges faster than algorithms that are more similar to the PageRank algorithm, as shown in Section 6.1.1.

4.2.3 CiteRank

The CiteRank algorithm, developed by Walker *et al.* [13], tries to overcome the problem of the aging effect in citation networks by taking the publication dates of papers into consideration. It is based on a similar idea as the PageRank algorithm, by simulating a random researcher that starts with a paper and follows citations until the researcher is satisfied with the search. At each point in the search, the researcher becomes satisfied and stops the search with a probability of α. Furthermore, CiteRank takes into consideration that a researcher usually starts investigating a research topic on recently published articles found in journals or conference proceedings and then continues following references to older publications.

Let $\boldsymbol{\rho}$ be the initial probability distribution, where the probability of selecting a paper i is $\rho_i = e^{-age(i)/\tau}$ which takes the age of a paper, $age(i)$, into consideration and defines τ to be the characteristic decay time.

Furthermore, let M be the transfer matrix containing the probabilities that a random researcher is following a citation. The matrix M is defined as follows: $M_{ij} = 1/\text{od}(p_j)$ if paper p_j cites paper p_i and zero otherwise. It follows that the probability of a researcher reaching a certain paper after following a single citation is given by $(1-\alpha)\cdot M\cdot\boldsymbol{\rho}$. Therefore, if a path of any length is allowed, the traffic is calculated using the following formula:

$$\boldsymbol{x} = I\cdot\boldsymbol{\rho} + (1-\alpha)\cdot M\cdot\boldsymbol{\rho} + (1-\alpha)^2\cdot M^2\cdot\boldsymbol{\rho} + \cdots \tag{4.2.10a}$$

$$= \frac{\boldsymbol{\rho}}{I-(1-\alpha)\cdot M} \tag{4.2.10b}$$

Using a different notation to describe the CiteRank algorithm, let $\boldsymbol{x}_0$ be the initial probability distribution $\boldsymbol{\rho}$, then for each iteration $1, 2, \ldots$ of the algorithm the CiteRank values can be computed with the following formula:

$$\boldsymbol{x}_t = \boldsymbol{x}_{t-1} + (1-\alpha)^t\cdot M^t\cdot\boldsymbol{\rho} \tag{4.2.11}$$

Similarly to the PageRank stopping criteria, the computation for the CiteRank algorithm stops when the result vector reaches the predefined precision threshold δ, namely $\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|_1 < \delta$. Note that the values of $\boldsymbol{x}_t$ are accumulative for each iteration and that the resulting vector has to be normalised such that $\|\boldsymbol{x_t}\|_1 = 1$ in order for the results to be comparable to the other algorithms.

In contrast to the PageRank approach of modeling random researchers that follow citations and restart their searches, the CiteRank algorithm does not compute the steady-state distribution of a Markov chain. Instead, the CiteRank algorithm rather models the dissemination of random researchers into the citation network until the change in the result vector falls below the precision threshold or all random researchers reach the outer edge of the citation network.

In addition to addressing P3, the initial distribution of random researchers onto the citation graph and the selection of the citation to follow depends on $\boldsymbol{\rho}$ and therefore addresses P1 and P2 since a random researcher is more likely to choose a recent paper when starting the search.

The drawback of CiteRank, compared to the PageRank algorithm, is that its time and space complexity is worse. Except for the first two iterations of CiteRank, matrix multiplication is required which generally has a time complexity of $O(n^3)$ where n is the number of vertices in a graph. Similarly, for solving Equation 4.2.10b, the computational

complexity of matrix inversion is also $O(n^3)$. It should be noted that M is very sparse for citation networks. Therefore, the computation using Equation 4.2.10a is faster than using Equation 4.2.10b. Furthermore, it is not guaranteed that the inverse of $I - (1-\alpha) \cdot M$ exists. The space requirement of CiteRank is $O(n^2)$.

4.2.4 NewRank

The NewRank algorithm [15] is a combination of both the PageRank and the CiteRank algorithm because it simulates the behaviour of random researchers using a Markov chain and incorporates the age of publications into the computation. Similarly to the CiteRank algorithm, let $\boldsymbol{\rho}$ be the vector containing the probabilities of selecting a paper, where $\rho_i = e^{-age(i)/\tau}$. As in the CiteRank algorithm, τ is the characteristic decay time and $age(i)$ is the age of paper i.

Let $D(p_i)$ be the probability of following a reference from paper p_i which is defined as follows:

$$D(p_i) = \frac{\rho_i}{\sum_{p_j \in N^+(p_i)} \rho_j} \tag{4.2.12}$$

The above equation simply normalizes the initial value of paper p_i by the initial values of all papers in its reference list. It follows from this equation that the likelihood of the random researcher following a young citation is greater than following a citation to a paper that is older.

The transition matrix A of the PageRank Markov chain from Equations 4.2.5 is updated such that it contains the elements $a_{ij} = \frac{D(p_i)}{\text{od}(p_i)}$. In addition, let $\boldsymbol{r}$ be the normalised vector such that $r_i = \frac{\rho_i}{\|\boldsymbol{\rho}\|_1}$. The initial probability distribution is then given by $\boldsymbol{x}_0 = \boldsymbol{r}$. For each iteration $i = 1, 2, \ldots$ the NewRank values are computed, similar to the PageRank algorithm, using the following formula:

$$\boldsymbol{x}_t = (1-\alpha) \cdot \boldsymbol{r} + \alpha \cdot \left(A^T + \boldsymbol{r} \cdot \boldsymbol{d}^T\right) \cdot \boldsymbol{x}_{t-1} \tag{4.2.13}$$

with the same stopping criteria as given in Equation 4.2.6.

Much like PageRank, the NewRank algorithm addresses P3, except that a random researcher is more likely to start a new search with a recently published paper, therefore also addressing problem P1. In addition, the random researcher is going to follow a citation to a more recent publication with a higher probability than choosing a citation that points to an older publication, addressing P2. This is shown in Section 4.2.6 with the graph in Figure 4.4. As with the PageRank algorithm, the NewRank score of a paper can be calculated using the Power Method. The time and space complexities of NewRank are both $O(n)$ per iteration which greatly improves on the requirements of the CiteRank algorithm.

4.2.5 Yet Another Paper Ranking Algorithm

In order to address problem P4 some metric that measures the prestige of publication venues has to be incorporated into the ranking algorithm. This was done by Hwang *et al.* [14] by proposing an algorithm that incorporates the Impact Factors of venues in their paper "Yet Another Paper Ranking Algorithm Advocating Recent Publications". In the following discussions this algorithm is referred to as YetRank (YR).

Similarly to CiteRank, let $\rho_i = \frac{1}{\tau} \cdot e^{-age(i)/\tau}$, where τ is the characteristic decay time and $age(i)$ is the age of the paper i. The impact factor of a venue v for a certain year y is

calculated by the Impact Factor method as described by Equation 4.1.2 with parameters: $IF(v, [y, y], [y-5, y-1])$.

Then the initial score for paper i published in the year y_i and at venue v_i is $s_i = IF(v_i, [y_i, y_i], [y_i - 5, y_i - 1]) \cdot \rho_i$. Furthermore, let $\boldsymbol{r}$ be the normalised vector such that $r_i = \frac{s_i}{\|\boldsymbol{s}\|_1}$.

As in the PageRank algorithm let A be the adjacency matrix where $a_{ij} = \frac{1}{\text{od}(p_i)}$ if paper i cites paper j and zero otherwise.

$$\boldsymbol{x}_t = (1-\alpha) \cdot \boldsymbol{r} + \alpha \cdot (A^T + \boldsymbol{r} \cdot \boldsymbol{d}^T) \cdot \boldsymbol{x}_{t-1} \tag{4.2.14}$$

By taking the impact factor of publishing venues into consideration, this algorithm addresses problems P1 through P5. The random researchers are more likely to start and restart their searches with papers that were published recently and in more renowned venues. This algorithm's time and space complexity is $O(n)$ for each iteration but requires an expensive once-off computation to compute the impact factors for each venue for each year.

4.2.6 Graph Examples

This section shows example graphs and corresponding results of the algorithms to demonstrate their behaviour and to point out some of the differences between them. The algorithms were initialised with the default parameters as stated by the authors in the papers in which the algorithms are defined. The following parameters were used with the precision threshold set to $\delta = 10^{-5}$.

CountRank (CCR): no parameters

PageRank (PR): $\alpha = 0.85$.

PageRank (PR2): $\alpha = 0.5$.

SceasRank (SR): $\alpha = 0.85$, $a = e$, $b = 1$.

SceasRank1 (SR1): $\alpha = 1$, $a = e$, $b = 1$, not adding edges to dangling vertices.

SceasRank2 (SR2): $\alpha = 0.85$, $a = e$, $b = 0$, not adding edges to dangling vertices.

CiteRank (CR): $\alpha = 0.31$, $\tau = 1.6$.

NewRank (NR): $\alpha = 0.85$, $\tau = 4.0$.

YetRank (YR): $\alpha = 0.85$, $\tau = 4.0$.

In this section, cells of tables are highlighted for better readability and indicate that they contain the largest values in a column.

As mentioned in Section 4.2.1, the PageRank algorithm unfairly favours vertices that exist within citation cycles. Figure 4.1 depicts a graph that contains two cycles[2]. The vertices 1 through 5 are all part of citation cycles and PageRank assign scores of 0.14 or more to each of them as shown in Table 4.1.

[2]The graphs in Figures 4.1, 4.2 and 4.3 are adapted from Sidiropoulos and Manolopoulos [11] who use the graphs to depict some of the drawbacks of the PageRank algorithm on bibliographic citation networks.

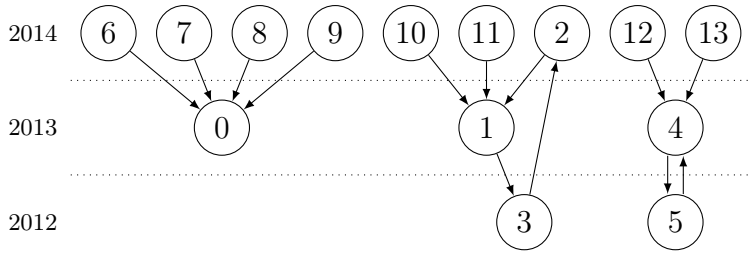


Figure 4.1: Illustrative Graph G_1.

Table 4.1: Ranking results for the graph G_1 in Figure 4.1.

Node	CCR	PR	PR2	SR	SR1	SR2	CR	NR	YR
0	0.31	0.06	0.12	0.22	0.22	0.12	0.11	0.07	0.07
1	0.23	0.16	0.13	0.19	0.20	0.11	0.14	0.16	0.16
2	0.08	0.14	0.09	0.09	0.10	0.08	0.11	0.14	0.14
3	0.08	0.15	0.10	0.12	0.13	0.09	0.11	0.15	0.15
4	0.23	0.19	0.13	0.20	0.21	0.11	0.14	0.18	0.18
5	0.08	0.17	0.11	0.12	0.13	0.09	0.11	0.16	0.16
6-13	0.00	0.01	0.04	0.01	0.00	0.05	0.03	0.02	0.02

Node 0 has the highest in-degree of 4 but only obtains a score of 0.06 according to PageRank. The same holds true for NewRank and YetRank since they are PageRank-like algorithms and therefore exhibit the same behaviour. If PageRank is computed with a damping factor of $\alpha = 0.5$ the advantage of being within citation cycles has a lesser effect, as seen in the fourth column in Table 4.1.

The column SR2 contains the ranking values of the SceasRank algorithm that models PageRank the closest.

Years were added to the graph in order to show the differences between CiteRank, NewRank and PageRank. If all vertices had the same year associated with them, then NewRank would be identical to PageRank. Similarly, YetRank's and NewRank's results are the same since all vertices were assigned the same impact factor of 1.

The graph in Figure 4.2 is used to demonstrate that PageRank transfers the weight of an important vertex to the vertices it cites. This is suitable for the Internet where a citation from an important website should bear more weight than the number of citations. In bibliometrics, this should still hold true but to a lesser degree since a single citation from an important paper should not outweigh the accreditation of many citations.

This is shown in Figure 4.2 where PageRank ranks vertex 0 higher than vertex 1 even though it has 5 fewer citations. NewRank and YetRank rank vertex 1 higher but only because of the publication dates that are taken into consideration. If all vertices had the same publication dates, then CiteRank would assign scores to vertices 0 and 1 of 0.29 and 0.33, respectively. Therefore, CiteRank does not transfer the weight of a vertices as freely to cited vertices compared to PageRank-like algorithms.

Similarly to the previous example, the balance between the weight of the number of citations and the weight of a single citation can be controlled by the damping factor for PageRank-like algorithms. This can be seen in the column PR2 where vertex 1 is ranked higher than vertex 0.

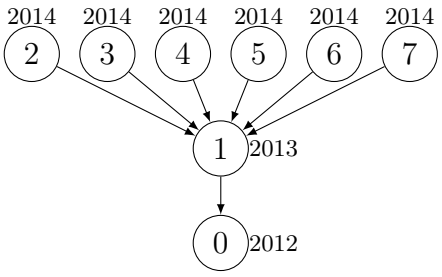


Figure 4.2: Illustrative Graph G_2.

Table 4.2: Ranking results for the graph G_2 in Figure 4.2.

Node	CCR	PR	PR2	SR	SR1	SR2	CR	NR/YR
0	0.14	0.34	0.23	0.30	0.35	0.18	0.25	0.32
1	0.86	0.33	0.31	0.60	0.65	0.27	0.33	0.34
2-7	0.00	0.05	0.08	0.02	0.00	0.09	0.07	0.06

Using the graph in Figure 4.3 Sidiropoulos and Manolopoulos [11] demonstrate that PageRank transfers weights easily along citation chains and that the effect of an important vertex is significant to the scores of vertices that are far down the citation chain. While their argument is true, they claim that the addition of vertex 8 to the graph increases the scores of vertices 4 and 5 by 6.82% and 7.14%, respectively. This is not accurate since the scores of vertices 4 and 5 actually decrease, once they are normalised by the number of vertices in the graph.

Table 4.3 shows the results of the algorithms when computing scores for the graph in Figure 4.3. The first number in each column is the score without vertex 8 added to the graph while the second number represents the score when vertex 8 is added. The results are normalised for SceasRank and CiteRank for comparison reasons and since the size of the graph changes.

Another important aspect about bibliographic citation networks is the publication dates of citing and cited papers. This plays an important role if the importance of a paper is coupled with its age. The graph in Figure 4.4 depicts a graph in which the vertices 0, 1, 2 and 3 are each cited three times. Comparing the ranking results of vertices 0 and 1 in Table 4.4 one can see that CiteRank, NewRank and YetRank assign a higher score to vertex 0 since it has a more recent date associated with it even though both vertices have exactly the same in-neighbourhood.

Similarly, vertices 2 and 3 both have an in-degree of three. However, vertex 2 receives three citation from vertices associated with 2014 while vertex 3's in-neighbourhood are vertices with dates of 2013. Vertex 2 should be ranked higher than vertex 3 since it is more often cited by vertices with more recent dates which can convey a higher relevancy if importance of a vertex is defined to be associated with current interest. From Table 4.4 one can see that vertex 2 receives higher scores than vertex 3 given by CiteRank, NewRank and YetRank.

Algorithms that do not consider dates cannot differentiate between the importance of the vertices 0 through 3.

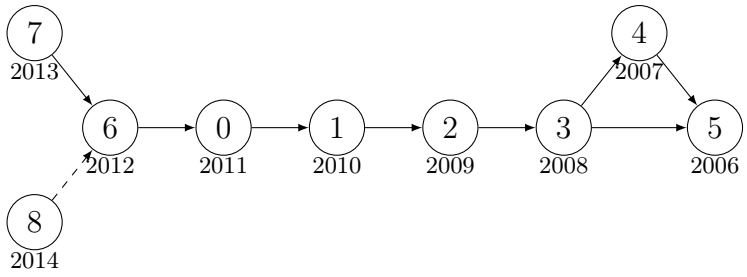


Figure 4.3: Illustrative Graph G_3.

Table 4.3: Ranking results for the graph G_3 in Figure 4.3.

Node	CCR	PR	SR1	CR	NR/YR
0	(0.13, 0.11)	(0.11, 0.11)	(0.14, 0.15)	(0.18, 0.17)	(0.14, 0.14)
1	(0.13, 0.11)	(0.13, 0.13)	(0.15, 0.14)	(0.15, 0.13)	(0.15, 0.14)
2	(0.13, 0.11)	(0.15, 0.15)	(0.16, 0.14)	(0.12, 0.10)	(0.15, 0.14)
3	(0.13, 0.11)	(0.17, 0.16)	(0.16, 0.14)	(0.09, 0.07)	(0.15, 0.13)
4	(0.13, 0.11)	(0.11, 0.10)	(0.08, 0.07)	(0.03, 0.03)	(0.09, 0.07)
5	(0.25, 0.22)	(0.21, 0.19)	(0.21, 0.18)	(0.06, 0.04)	(0.14, 0.12)
6	(0.13, 0.22)	(0.08, 0.09)	(0.10, 0.17)	(0.20, 0.21)	(0.11, 0.13)
7	(0.00, 0.00)	(0.04, 0.03)	(0.00, 0.00)	(0.16, 0.08)	(0.07, 0.05)
8	(–, 0.00)	(–, 0.03)	(–, 0.00)	(–, 0.16)	(–, 0.06)

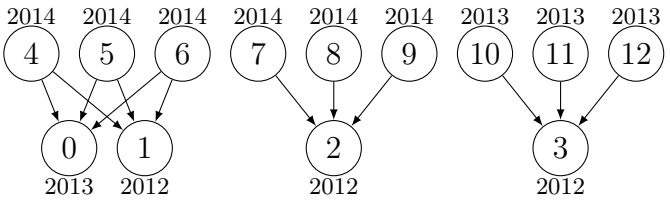


Figure 4.4: Illustrative Graph G_4.

Table 4.4: Ranking results for the graph G_4 in Figure 4.4.

Node	CCR	PR	SR	SR1	SR2	CR	NR	YR
0	0.25	0.13	0.20	0.25	0.12	0.11	0.12	0.11
1	0.25	0.13	0.20	0.25	0.12	0.09	0.10	0.10
2	0.25	0.13	0.20	0.25	0.12	0.17	0.18	0.18
3	0.25	0.13	0.20	0.25	0.12	0.10	0.14	0.14
4-9	0.00	0.08	0.03	0.00	0.08	0.07	0.06	0.06
10-12	0.00	0.08	0.03	0.00	0.08	0.04	0.04	0.04

4.3 Venue Ranking Algorithms

This section shows how methods described in the previous section can be adapted to rank publication venues such as journals, conferences, authors or academic institutions instead of individual papers.

The simplest approach is to use one of the algorithms described in the previous section and to compute the average score of the papers associated with venues. Let $\mathcal{V}$ be the set of venues where $\mathcal{P}(v)$ is the set of papers associated with venue v. Given the PageRank scores $PR(p)$ for all papers p in a citation network G, then for each venue $v \in \mathcal{V}$ a ranking score PRV is computed by the following formula:

$$PRV(v) = \frac{\sum_{p \in \mathcal{P}(v)} PR(p)}{|\mathcal{P}(v)|} \tag{4.3.1}$$

Using this approach can lead to unfair advantages of smaller publication venues with a small number of papers with high paper scores. For example, assume venue A has two papers with scores 10 and 1. Furthermore, let venue B have 20 papers of which 5 have scores of 10 and the others have scores of 1. Venue A would have an average result score of 5.5 while venue B's score would be 3. It is reasonable to assume that venue B should be ranked higher than venue A since B has five times the number of high impact papers than venue A.

Alternatively, PageRank can be computed over a journal cross-citation graph which the Eigenfactor Metric [66] does and is further discussed in Section 4.3.1. Similarly an author co-citation graph can be constructed from a bibliometric citation network and used with PageRank. This was done by West *et al.* [10] and is formulated in Section 4.3.2.

4.3.1 The Eigenfactor Metric

The Eigenfactor calculation is also based on the PageRank algorithm where random researchers traversing a graph are modelled using a Markov chain. Instead of the underlying graph consisting of papers and citations, papers published at the same journal or conference are aggregated into a single vertex and edges between these vertices indicate the number of references between these subsets of papers.

Let G_J be this aggregated graph representing the journal cross-citations. The vertices in the graphs are distinct venues and weighted directed edges between journals indicate the number of citations from one to another journal.

However, not all citations between venues are included into the graph. Similar to the Impact Factor, the Eigenfactor metric incorporates two time frames. The census window is the current year for which the Eigenfactor rankings are computed while the previous 5 years constitute the target window. For example, if the journal graph was to be constructed for computing rankings for the year 2013, then only references from papers published in 2013 to papers published in the years 2008 through 2012 would be included. This is done in order to compute current importance values and not overall rankings for the venues.

Let A be the normalised adjacency matrix corresponding to the graph G_J whose elements are computed as follows:

$$A_{ij} = \frac{w_{ij}}{\sum_{k \in N^+_{G_J}(i)} w_{ik}} \tag{4.3.2}$$

The element A_{ij} is the number of citations from articles in the census window and published in journal i that reference articles published within the target window and in journal j, normalised by the total number of outgoing references of journal i. If no such citations exist, the element A_{ij} is zero. Furthermore, since all self-citations are ignored in the Eigenfactor method, all diagonal entries in A are zero as well. In [66] Bergstrom and West state the reasoning behind their decision to exclude journal self-citations. Firstly, they want to discourage opportunistic self-citation practices of journals which can lead to increased ranking scores. Secondly, they argue that small journals with unusual citation patterns might appear as nearly-dangling due to a high percentage of self-citations which would unfairly increase their overall score.

Let $\mathcal{P}(i)$ be the set of papers published by journal i. Then the vector $\boldsymbol{r}$ contains the number of papers published by a journal during the time frame of the target window, normalised by the total number of papers in the graph, for each journal. Or more concisely, $r_i = |\mathcal{P}(i)|/n(G_J)$.

The random researchers are evenly distributed initially (i.e. $\boldsymbol{x}_0 = 1/n(G_J)$) and solving for the leading eigenvector, each iteration $i = 1, 2, \ldots$ of the power method is computed as follows

$$\boldsymbol{x}_t = (1-\alpha)\cdot\boldsymbol{r} + \alpha\cdot\left(A^T + \boldsymbol{r}\cdot\boldsymbol{d}^T\right)\cdot\boldsymbol{x}_{t-1} \tag{4.3.3}$$

until a predefined precision threshold δ is reached which results in the approximation of the steady-state distribution $\boldsymbol{\pi}$ of the corresponding Markov chain.

The Eigenfactor scores are then computed according to the following equation:

$$EF = 100\cdot\frac{A^T\cdot\boldsymbol{\pi}}{\|A^T\cdot\boldsymbol{\pi}\|_1} \tag{4.3.4}$$

which results in a score between 0 and 100 denoting a journal's overall influence.

The Eigenfactor metric also computes an *Article Influence Score* (AI_i) for each journal i which represents a per-article influence of a journal and is calculated as follows

$$AI_i = 0.01\cdot\frac{EF_i}{r_i} \tag{4.3.5}$$

The AI scores for journals can be used to compare against their Impact Factor values.

It may be noted that the restart of the random researchers in Equation 4.3.3 is not evenly distributed over the journal graph but is weighted by the vector $\boldsymbol{r}$ which contains values that are proportional to the article counts of journals. Therefore, the probability that a random researcher selects a large journal is higher than for a journal that contains a small number of articles. This is to ensure that rankings of smaller journals are not unfairly inflated. When the construction of the journal cross-citation graph is ignored, the time and space complexity of the Eigenfactor metric is $O(n)$ per iteration where n is the number of journals in the citation network.

4.3.2 The Author-Level Eigenfactor Metric

In [10], West *et al.* demonstrate how to apply the Eigenfactor metric to author co-citation graphs. The Eigenfactor metric is simply the PageRank algorithm applied to a normalised author co-citation graph that is constructed from a data set that contains information about authors in addition to articles and references.

Let G_C be a bibliographic citation graph and $\mathcal{A}$ be the set of authors, where $\mathcal{A}(p_i)$ is the set of authors that authored paper p_i. Similarly, let $\mathcal{P}(a_i)$ be the set of papers authored

by author a_i. The author co-citation graph G_A, used as input for the Author-Level Eigenfactor method, is then constructed as follows:

Step 1 - Normalising the citation network G_C:

$$w_{G_C}(p_i, p_j) = \frac{1}{|\mathcal{A}(p_i)| \cdot |\mathcal{A}(p_j)| \cdot \mathrm{od}_{G_C}(p_i)} \tag{4.3.6}$$

The equation above normalises the weight of an edge (p_i, p_j) by the product of the number of authors in the citing paper p_i, the number of authors in the cited paper p_j, and the number of references in the bibliography of paper p_i.

Equation 4.3.6 divides the credit of an incoming citation equally between the co-authors of a paper because the average sizes of collaboration groups differ between various academic disciplines. Otherwise, authors that commonly work in larger groups of co-authors would be unfairly advantaged because they would receive full accreditation of a citation.

Step 2 - Constructing the author co-citation graph G_A:

$$w_{G_A}(a_i, a_j) = \sum_{\{(p_i,p_j) \in E(G_C) | p_i \in \mathcal{P}(a_i) \wedge p_j \in \mathcal{P}(a_j)\}} w_{G_C}(p_i, p_j) \tag{4.3.7}$$

The author co-citation graph is constructed by inserting edges $w_{ij} = (a_i, a_j)$ whose weights correspond to the sum of the edges from the citation network G_C of papers p_i associated with author a_i that cite papers p_j written or co-authored by author a_j.

Step 3 - Normalizing the adjacency matrix $A(G_A)$:

$$\begin{aligned} A_{ij} &= \frac{w_{G_A}(i,j)}{\sum_{k \in N^+_{G_A}(i)} w_{G_A}(i,k)} && \forall i \neq j \\ A_{ij} &= 0 && \forall i = j \end{aligned} \tag{4.3.8}$$

The above equation ensures that A is a stochastic transition matrix for the Markov process. The diagonal values are set to zero so that author self-citations are omitted. For multi-authored papers, this step only removes the citation credit for the authors who are self-citing. The citation is still counted for authors that only co-authored either the cited article or the citing article.[3]

Let the vector $\boldsymbol{r}$ contain the number of articles written by each author normalised by the total number of articles in the graph. Formally, let $r_{a_i} = |\mathcal{P}(a_i)|/n(G_C)$.

For completeness the equation of the Eigenfactor metric for the power iteration is given again below which is the same as in Equation 4.3.3.

$$\boldsymbol{x}_t = (1-\alpha) \cdot \boldsymbol{r} + \alpha \cdot \left(A^T + \boldsymbol{r} \cdot \boldsymbol{d}^T\right) \cdot \boldsymbol{x}_{t-1} \tag{4.3.9}$$

Again, the above sequence $\boldsymbol{x}_t$ converges to the eigenvector $\boldsymbol{\pi}$ corresponding to the principal eigenvalue. The computation of the power iteration stops when the precision $\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|_1 < \delta$ is reached.

[3]Setting the diagonal values to zero can lead to the occurrence of zero rows or zero columns which, respectively, indicates that an author either only cited his own single-authored work or his articles were only cited by single-authored papers that were published by the same author. In these cases the associated author can simply be removed from the graph.

From Equation 4.3.9 one may notice that the probabilities related to the restarts of the random researcher are weighted by $\boldsymbol{r}$, which contains values proportional to the number of articles written by an author. This is required to ensure that the random restarts do not favour authors with only a few articles published.

To compensate for the bias that is introduced by the restarts of the random researchers that favour authors that are rarely cited, the result scores of the eigenvector $\boldsymbol{\pi}$ are weighted by the normalised incoming citations for that author. The final Author-Level Eigenfactor (AF) ranking scores are therefore computed as follows:

$$AF = 100 \cdot \frac{A^T \cdot \boldsymbol{\pi}}{\|A^T \cdot \boldsymbol{\pi}\|_1} \tag{4.3.10}$$

The above equation computes scores for authors between 0 and 100 and can be interpreted as the overall impact or importance of an author. The Author-Level Eigenfactor method has a time and space complexity of $O(n)$ where n is the number of authors in the citation network.

4.3.3 Graph Example

The graph in Figure 4.5 depicts a citation network where the vertices represent papers. Each paper is associated with a distinct venue v_0 through v_4 as indicated on the left hand side of the graph and authors a_i labelled next to each vertex.

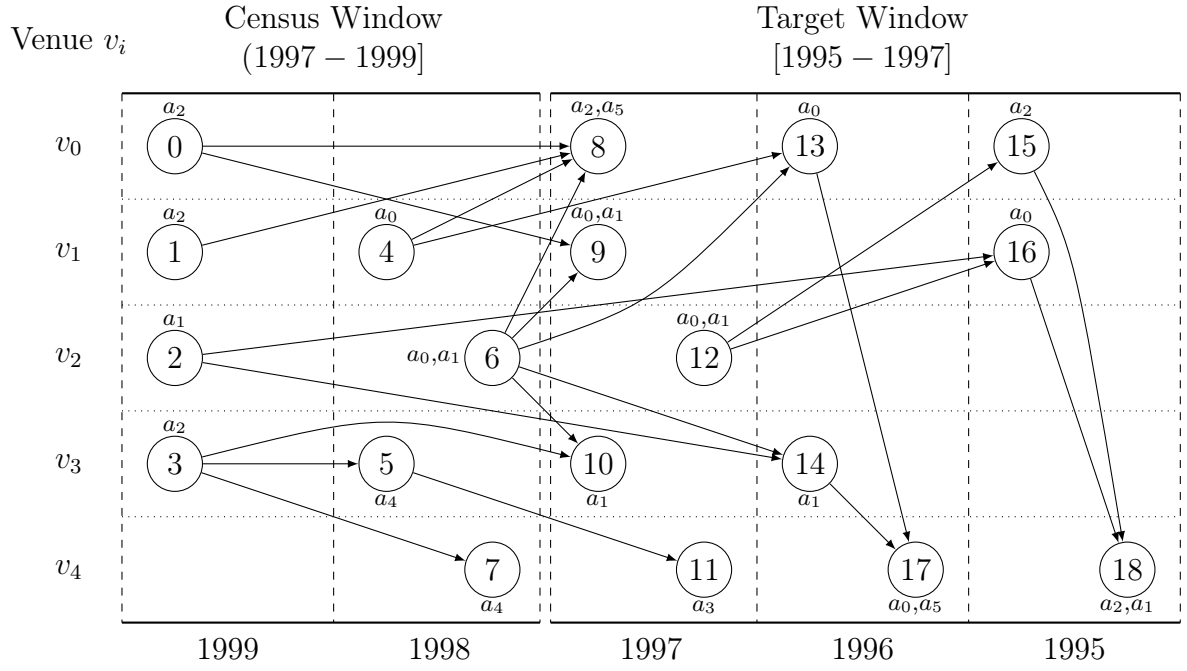


Figure 4.5: Illustrative Graph G_5. Each vertex represents a paper that is associated with a year, a venue v_i, and a set of authors a_i.

In addition, publication dates of papers are given at the bottom of the graph. The census and the target window, displayed at the top, are used by the Eigenfactor metric and the Impact Factor method which uses a venue-cross citation graph that is constructed only from citations that originate from the papers in the census window and cite papers that are within the target window.

Note that papers of v_2 are never cited. This can be seen in the resulting venue cross-citation graph which is depicted on the left in Figure 4.6. Here, the in-degree of v_2 is zero. The graph is constructed by considering the census and target windows. Observe that the weight of the edge (v_3, v_4) is only 1 because the only citation that originates from the census window and ends in the target window is the citation from paper 5 referencing paper 11.

The author co-citation graph extracted from graph G_5 is shown on the right in Figure 4.6. The graph includes author self-citations and a single citation is counted multiple times where more than one author is associated with either the citing or cited paper. For example, the citation from 6 to 9 is counted 4 times because both papers are authored by authors a_0 and a_1.

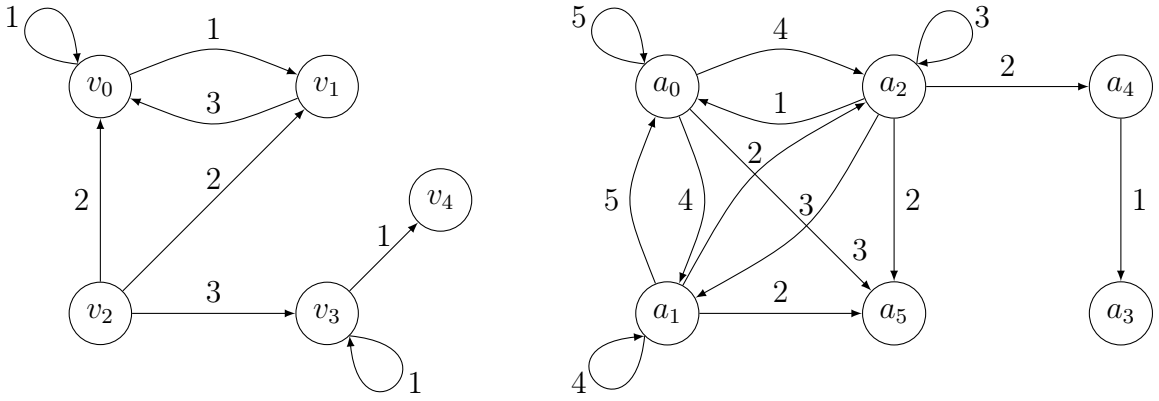


Figure 4.6: Illustrative Graph G_6. On the left the venue cross-citation graph extracted from G_5 is depicted. Similarly, the author co-citation graph associated with G_5 is shown on the right.

Table 4.5 shows the results of the Eigenfactor method (EF) and the corresponding Article Influence (AI) scores of the journal cross-citation graph G_6. It also lists the journal CountRank values (CCR), the results of the Impact Factor (IF) method and the average PageRank score per venue (PRV). Both the Eigenfactor and Impact Factor methods were used with the same census and target windows as shown in Figure 4.5. The damping factor for the Eigenfactor method was set to $\alpha = 0.85$. All results are normalised for easier comparison.

Table 4.5: Ranking results of the venue cross-citation graph in Figure 4.6. For easier comparison the Eigenfactor (EF) scores are normalised to sum up to 1.

Node	CCR	EF	AI	IF	PRV
v_0	0.42	0.46	0.37	0.34	0.21
v_1	0.25	0.47	0.57	0.36	0.17
v_2	0.00	0.00	0.00	0.00	0.12
v_3	0.25	0.01	0.01	0.34	0.16
v_4	0.08	0.06	0.05	0.06	0.33

Note that CountRank, the Eigenfactor method, and the Article Influence scores do not take venue self-citations into consideration. However, the Impact Factor method is

defined to include self-citations. The PageRank values for papers are computed without regarding venue information and therefore venue self-citations are intrinsically included.

Table 4.6 shows the author citation counts (CCR), the output of the Author-Level Eigenfactor method (AF), as well as the average PageRank scores per author of the author co-citation graph in Figure 4.5.

Table 4.6: Ranking results of the author co-citation graph in Figure 4.6. For easier comparison the Author-Level Eigenfactor (AF) scores are normalised to sum up to 1.

Node	CCR	AF	PRA
a_0	11	0.22	0.14
a_1	11	0.20	0.14
a_2	9	0.17	0.16
a_3	1	0.10	0.18
a_4	2	0.06	0.11
a_5	7	0.25	0.28

Note that the Author-Level Eigenfactor method is the only method that does not include author self-citations.

4.4 Chapter Summary

This section formally defined metrics that are commonly used in bibliometrics such as the h-index and the Impact Factor, and described ranking algorithms that have recently been introduced in the literature. The ranking algorithms that are PageRank-like or model traffic-flow through a citation network are defined mathematically. For each metric the theoretical advantages and drawbacks are highlighted and discussed. In addition, illustrative graphs depict how the methods are used on citation networks to rank papers, authors and venues.

Chapter 5

Data Sets

For the experiments and analyses in this thesis, citation networks are constructed from two different data sets and used as input for the ranking algorithms.

Firstly, a data set assembled by Tang *et al.* [6] who extract citation information from the DBLP database is used. This data set mainly contains academic papers from the Computer Science domain. Using this data set a citation network with 469 940 vertices and 2 083 983 edges is constructed as described in Section 5.1.

Secondly, a data set from Microsoft Academic Search (MAS) that contains information about 39 million academic articles and over 262 million references. This data set contains papers from various academic disciplines such as Computer Science, Chemistry, and the Arts and Humanities. More information on the MAS data set and how it is used in this thesis is given in Section 5.2.

For evaluation purposes, further data sets that are based on expert opinions are used. These data sets were collected by hand and contain, for example, papers that won best paper awards at conferences or authors that received accolades due to their innovative and continuing contributions to their fields of research. These data sets are further described in Section 5.3.

5.1 DBLP Data Set

The DBLP Computer Science Bibliography is a database hosted at Universität Trier [7] and tracks the most important journals and conference proceedings in the Computer Science (CS) domain. A data set[1] published by Tang *et al.* [6] which contains data extracted from the DBLP database and citation information obtained from the ACM Digital Library [67] is used in this thesis and referred to as the "DBLP" data set in the following chapters.

The DBLP data set contains 2 244 018 papers, 2 083 983 references and 8 867 venues. All papers within this data set are associated with a year. The citation network constructed from this data set contains 469 940 vertices and 2 083 983 edges since vertices with a degree of 0 and papers not associated with a venue are removed from the data set.

This results in 4.43 references per paper which is relatively small. Considering only papers with an in-degree of one or more, the average in-degree is 6.53. Similarly, the average out-degree of non-dangling vertices is 6.64. In other words, this network contains 162 895 internal papers (34.66%) that contain at least one reference and are cited by

[1] The source data is available freely at `http://arnetminer.org/citation`.

Table 5.1: Properties of the DBLP data set and the associated citation network constructed from this data set. The citation network contains 469 940 vertices and 2 244 018 edges with vertices having an average in-degree of 6.53 if papers with no incoming citations are ignored.

Description	Property
Papers	2 244 018
Venues	8 867
Graph Order	469 940
Graph Size	2 083 983
Vertices with $\mathrm{id}(n) > 0$ (V_I)	319 210
Vertices with $\mathrm{od}(n) > 0$ (V_O)	313 625
$V_I \cap V_O$	162 895
Avg. In-Degree	6.53
Avg. Out-Degree	6.64

another paper at least once. In the DBLP citation network there are 156 315 dangling vertices (33.26%) and 150 730 (32.07%) vertices that have an in-degree of zero.

Simple string comparison is used to match venues and it should be noted that no author name disambiguation was performed on this data set. Therefore, this cleaned up citation network is used in the following chapters for experiments that do not require author information.

In order to assess the quality of the data set, a random sample of 10 papers was selected from the 469 940 papers in the citation network. These 10 papers contain 219 references in their reference lists of which 101 papers (46.12%) were found in the DBLP data set at hand.

The number of papers from the reference lists that are matched to entries in the DBLP data set is very low and less than half are found. This low figure can be partially attributed to a large number of references to papers that fall outside the scope of the DBLP data set since they reference papers published at venues that cover other academic disciplines not indexed by DBLP. It is difficult to determine which papers fall outside DBLP's scope by simply looking at the papers' venues and deciding whether the referenced papers should be indexed by DBLP or not. Therefore, all referenced papers are considered.

In order to obtain a coverage value for the number of citations in the DBLP data set, the number of papers that are indexed and can be referenced has to be found. After categorising the 219 references to exclude references to webpages, technical report and lecture notes and only including journal articles, conference proceedings and books (including PhD and Masters theses), 183 references were counted. This results in 55.19% of referenced papers found in the DBLP data set.

The sum of the out-degree of the 10 sample papers in the citation network is 29 which results in 28.71% of the 101 references being identified in the DBLP data set. To compute an accuracy value of the edges of the DBLP citation network, the same set of 10 papers was used and their references in the data set checked against the entries in their reference lists. All of the 29 references in the DBLP citation network point to the correct paper, yielding a 100% accuracy.

Therefore, in terms of citations, the citation network constructed from the DBLP data set has a low coverage (28.71%) and a high precision (100%) based on the references found in the 10 sample papers.

Lastly it should be noted that 105 papers (47.95%) were found by searching the official DBLP website compared to the 101 papers (46.12%) that were found in the DBLP data

set used in this thesis. The difference between the two data set is very small with 1.83% more papers found through the official DBLP website.

The overall quality of the DBLP data set is relatively low. Therefore, not all experiments use the DBLP citation network and it is only used for comparison reasons against the Microsoft Academic Search data set which is described in the following section.

5.2 Microsoft Academic Search Data Set

Microsoft Academic Search is a search engine for academic papers developed by Microsoft Research. The data set extracted from this service's indexed data is referred to as the MAS data set in the following sections.[2] The source data set is an integration of various publishing sources such as Springer and ACM.

The entities that are extracted from the data set and processed for the experiments and analyses in the following sections are papers, authors, publication venues and references. The raw count of these entities are as follows; 39 846 004 papers, 19 825 806 authors and 262 555 262 references. Furthermore, it includes information about 21 994 journals and 5 190 conferences.

Publication venues and each paper published there are assigned to exactly one domain. For example, all papers published at the *International Conference on Software Engineering* (ICSE) are associated with the CS domain. This property is useful when comparing publication trends between different academic domains and for analysing the effect that cross-domain references have on the results of the various ranking algorithms. Table 5.2 lists the individual domains and the total number of papers that are assigned to each domain.

Table 5.2: Paper counts per domain in the MAS data set. The column "Paper Count" displays the number of papers that have a venue and a publication year associated with them. The last column indicates the number of "bad papers" which cannot be used for the experiments since they either are not associated with a venue or do not contain a publication year.

Domain	Raw Paper Count	Paper Count	Bad Papers
Agriculture Science	457 677	454 898	0.61%
Arts & Humanities	1 351 369	1 349 267	0.16%
Biology	3 670 904	3 649 683	0.58%
Chemistry	4 186 521	4 171 812	0.35%
Computer Science	2 280 595	2 245 652	1.53%
Economics & Business	823 953	817 728	0.76%
Engineering	2 062 348	2 044 874	0.85%
Environmental Sciences	426 687	422 860	0.90%
Geosciences	745 814	740 894	0.66%
Material Science	848 257	843 003	0.62%
Mathematics	995 139	989 445	0.57%
Medicine	11 164 334	11 097 164	0.60%
Physics	2 007 333	2 001 171	0.31%
Social Science	1 729 693	1 725 187	0.26%

[2]The database from Microsoft Academic Search was received in October 2013 and is now available at `https://datamarket.azure.com/dataset/mrc/microsoftacademic`.

Note that about 20.58% (8 202 242) of all papers do not have a publication venue and are therefore not associated with a specific domain. These papers are not included in the raw numbers in Table 5.2 and are also excluded from any experiments and analyses.

It is important to obtain a uniform data set for the comparability of results. Therefore, the raw data as described previously had to be cleaned up in order to construct a consistent citation network for the various experiments and analyses. For example, papers need to have a publication year associated with them in order to include them in time series analyses. Some algorithms that were described earlier depend on the venue at which articles are published, therefore requiring papers to be assigned to a distinct journal or conference.

Consider, for example, the 2 280 595 papers in the CS domain, as shown in Table 5.2. Of these papers, 34 943 (1.53%) are bad papers that do not have a publication year associated with them and are therefore excluded when constructing the citation network.

Table 5.3: The number of references per domain in the MAS data set. The references are displayed according to their type. For example, the column "Dest. in Set" indicates the number of references that originate from a non-domain paper and reference a domain paper. Similarly, the column "Src. in Set" shows the number of references in a domain that originate from a paper in the domain and reference a paper that falls outside of the domain. The column "Internal" lists the number of citations that both originate from and terminate at papers that belong to the associated domain.

Domain	Dest. in Set	Internal	Src. in Set	% Internal
Agriculture Science	1 488 599	1 046 817	1 785 710	24.23%
Arts & Humanities	836 049	345 814	928 098	16.39%
Biology	18 066 426	19 763 110	19 900 923	34.23%
Chemistry	12 855 729	7 598 217	8 197 453	26.52%
Computer Science	10 819 149	10 691 968	8 832 749	35.24%
Economics & Business	4 008 532	2 845 259	3 159 460	28.41%
Engineering	4 873 754	2 571 933	5 305 940	20.17%
Environmental Sciences	2 483 011	906 302	2 568 810	15.21%
Geosciences	3 576 993	2 873 503	3 383 607	29.22%
Material Science	1 472 364	1 058 363	1 837 898	24.23%
Mathematics	3 274 078	2 136 353	2 323 825	27.62%
Medicine	25 318 683	55 379 728	22 181 131	53.83%
Physics	4 269 665	1 944 779	4 610 285	17.97%
Social Science	3 929 363	2 274 789	4 204 296	21.86%

When selecting references for constructing a citation network from a subset of the data set such as the CS domain, certain properties have to be taken into consideration. Since research is conducted across domains, all references from papers that fall outside of the CS domain and cite a paper within the domain have to be added to the graph. Similarly, references from CS papers that cite non-CS papers have to be added to the citation network too for certain experiments. For example, if the average reference age of references from CS papers is calculated, all outgoing references have to be considered and therefore should be included in the analysis.

Table 5.3 list the total number of references for each domain, categorised into the three different reference types. For example, considering only the CS papers, there are 10 691 968 internal citations, which are references that originate from CS papers and cite papers that also fall within the CS domain. 8 832 749 references originate from CS papers

and reference papers of other domains. Similarly, 10 819 149 references are contained in the data set whose destinations are CS papers but that originate from papers outside the CS domain. Therefore, 35.24% of references of the constructed network are domain internal references.

Table 5.4: The size of the cleaned MAS data set. The number of papers and references for each domain are listed. In addition, the number of vertices and edges of the citation networks constructed from this data are shown in the columns "Graph Order" and "Graph Size", respectively.

Domain	Total Papers	References	Graph Order	Graph Size
Agriculture Science	814 371	2 157 687	669 802	2 043 049
Arts & Humanities	1 589 649	836 289	489 266	808 746
Biology	5 812 446	32 570 153	5 050 140	31 109 998
Chemistry	6 347 301	18 533 656	4 818 025	17 497 374
Computer Science	3 066 801	16 046 156	2 394 976	12 907 440
Economics & Business	1 216 855	4 430 340	897 516	4 195 391
Engineering	2 995 970	6 594 726	2 054 552	5 148 489
Environmental Sciences	972 911	2 868 126	852 752	2 618 201
Geosciences	1 119 648	5 780 819	864 291	4 255 434
Material Science	1 189 381	2 369 525	833 169	2 144 649
Mathematics	1 642 396	4 256 927	1 290 464	3 617 518
Medicine	12 785 698	68 813 364	8 713 225	66 024 896
Physics	2 754 641	5 510 968	1 749 451	3 922 099
Social Science	2 482 765	4 455 332	1 547 746	4 300 297

When constructing citation networks for the computation of the PageRank and similar algorithms, only the references that point to CS papers are required (see Section 2.3). Therefore, the resulting network consists of 3 066 801 papers (adding 786 206 non-domain papers) and 16 046 156 references (removing references that originate from non-domain papers that are bad).

The final citation network for the CS domain consists of 2 394 976 vertices and 12 907 440 edges. The lower count of papers is due to the fact that some papers contain invalid year values such as -1 or 2050 but mostly because they are isolated vertices that have neither incoming nor outgoing edges. Furthermore, 3 138 716 references were removed because the papers where they originate or terminate are bad. This process is used for all domains and the resulting citation network properties are given in Table 5.4.

For the evaluation of the ranking algorithms, this cleaned up CS citation network is used because of the nature of the evaluation data which are papers and authors from the CS domain.

5.3 Evaluation Data Sets

Four different types of test data sets that are based on expert opinions are used for the experiments in this thesis. The entries in these data sets were collected by hand and are described in further detail in the following sections.

5.3.1 High-Impact Paper Awards

A data set of high-impact papers, often called most influential papers (MIP), was compiled for different CS conferences. A most influential paper is an accolade awarded to papers post-publication, usually 10 to 15 years after the initial publication of the paper.

The prize signifies that a paper has had the most impact over the intervening years in terms of research, methodology or application. Conferences that hand out these types of awards are predominantly in the CS domain with varying guidelines on the selection processes, but the prizes signify the same meaning of influence and impact.

Usually a single paper is awarded this prize at a conference in a given year but it does occur that two or more papers tie in the selection process and therefore more than one MIP prize is awarded in a year at some conferences. In total 210 papers were found from 14 different venues and matched against the MAS and DBLP data sets. Of these, 207 papers are contained in the MAS data set while 151 of the papers could be matched against entries in the DBLP data set. A list of the conferences that hand out this type of award and were selected for this test data set, is given in Table A.2 in Appendix A.2.

These papers are referred to as **award papers** in the following chapters. This data set of award papers is used to measure the accuracy of the algorithms in identifying and ranking high-impact papers. The results of this analysis are presented in Section 7.1.

5.3.2 Best Paper Awards

The second type of data that was collected contains articles that were awarded the prize of best paper at a conference in the year that they were published. At conferences this prize is usually awarded to one or more articles that are considered to be of the highest quality in the given year by a review panel. In the following discussions these papers are referred to as **best papers**. In total 464 papers from 32 different venues were collected and matched to the corresponding entries in the MAS data set. These papers are used to evaluate the 32 venues on how well they predict high-impact papers. The results of this experiment are given in Section 7.2.

5.3.3 Author Contribution Awards

In order to assess the performance of the venue ranking algorithms, test data that contains authors or journals is required. For this purpose 19 lists of in total 268 researchers that won an award for their innovative, highly significant and enduring contributions to their fields were collected. Of the 268 prize recipients, 18 authors have won two or more prizes. In total 249 distinct authors were matched to corresponding entries in the MAS data set. This set of authors is referred to as **award authors** in the following chapters. A detailed description of the awards handed out at various conferences is given in Table A.4 in Appendix A.2. The results of evaluating venue ranking algorithms using this data set are given in Section 7.3.

5.3.4 Important Papers

Lastly, a list of **important papers** in the CS domains was compiled. The source for this list is Wikipedia [68] where papers that are regarded important to a research field were selected by Wikipedia editors. According to the guidelines on the Wikipedia webpages themselves, an important paper can be any type of academic publication given that it

meets at least one of the following three conditions. Firstly, a publication that led to a significant, new avenue of research in the domain in which it was published. Alternatively, a paper is regarded as a breakthrough publication if it changed the scientific knowledge significantly and is therefore judged noteworthy enough to be granted a place on this list. Thirdly, influential papers that changed the world or had a substantial impact on the teaching of the domain, are also included in the list of important papers. From the papers listed on Wikipedia 115 were matched against paper entries in the MAS data set that contain venue and publication year information. This data set is used to evaluate how well the various ranking algorithms can identify these important papers. The results of this experiment are discussed in Section 7.4.

5.4 MAS Data Set Properties

In this section publication trends on the MAS data set are depicted. The MAS data is partitioned into broad academic disciplines such as Mathematics and Computer Science.

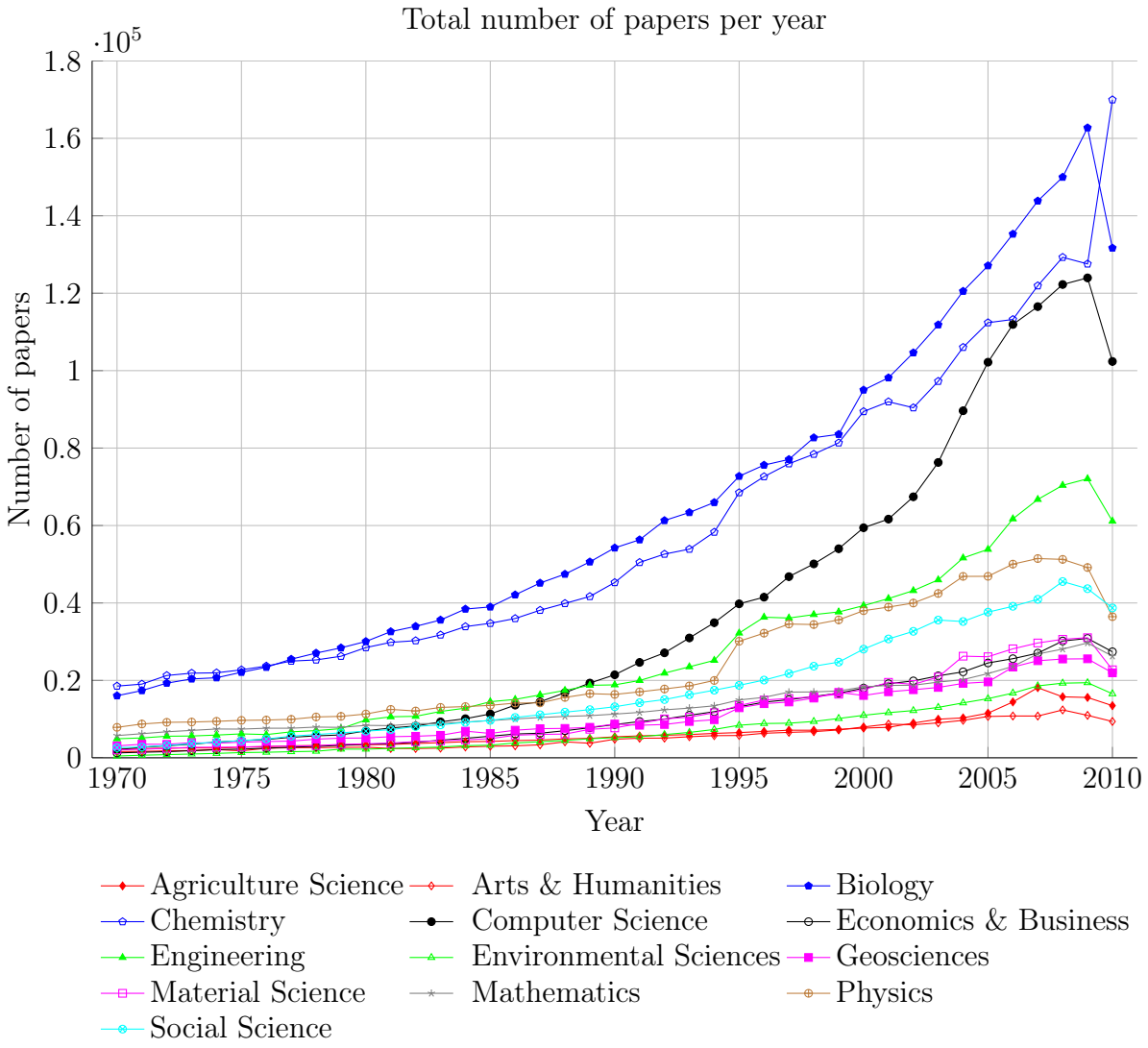


Figure 5.1: The total number of papers produced in the different domains over time.

These partitions are used to identify publication trends that differ between academic domains. It should be noted that the following analyses are merely indications of publication trends and cannot be seen as definitive results. Nonetheless, some insights into the properties of the MAS data set can be obtained and are discussed in this section.

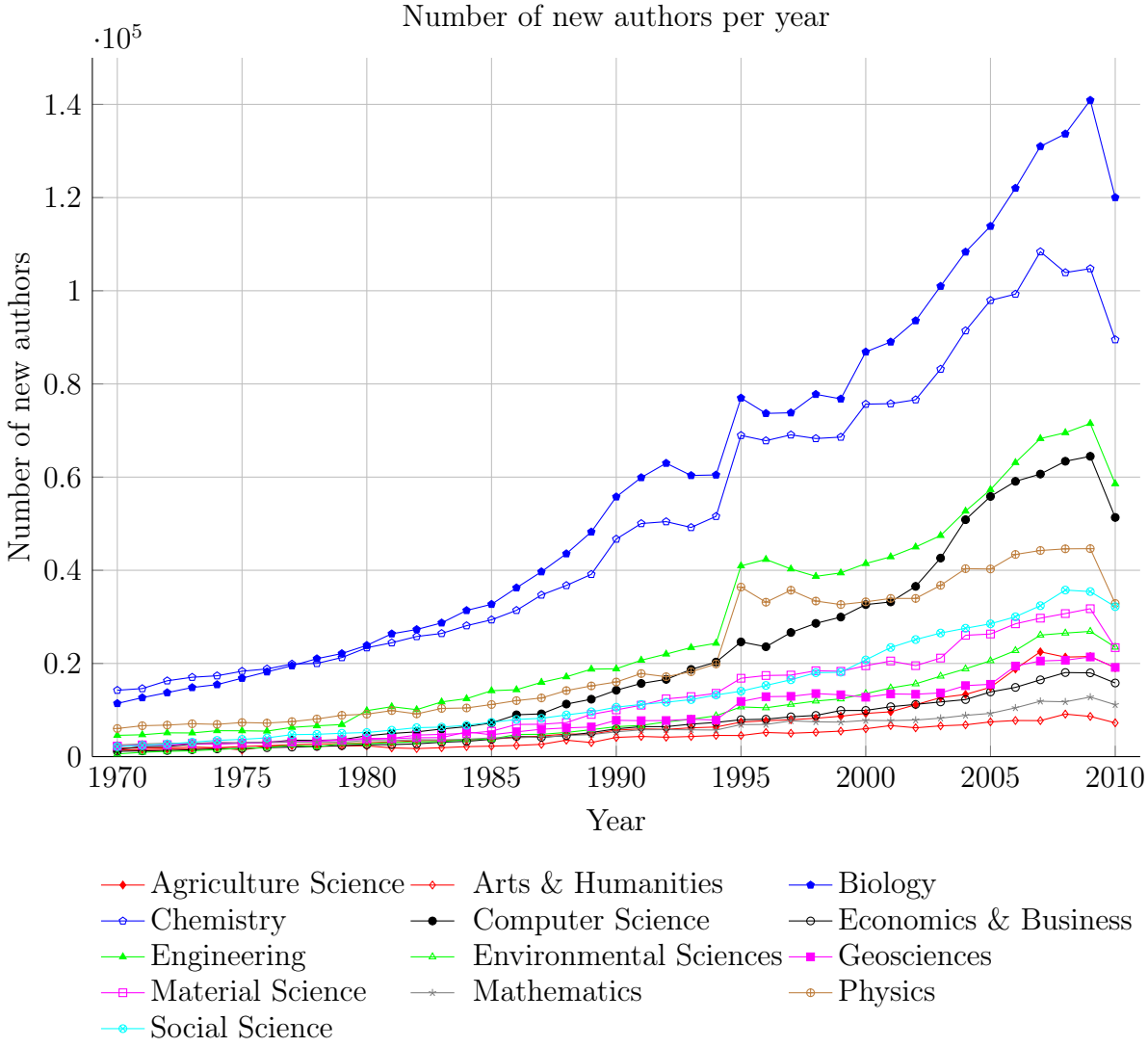


Figure 5.2: The number of new authors that publish their first publications over time.

As discussed before, the venue at which papers are published determines the discipline into which papers are categorised. Some publishing venues, such as Nature or Science, are multi-disciplinary and cannot easily be categorised into a single discipline. Therefore, the number of papers that are published over the years (Figure 5.1) cannot be seen as the size of the respective disciplines. Furthermore, it is difficult to reason about the sizes of the disciplines because the data for MAS is collected from various publishers and online sources and is not exhaustive. The data is then combined, sorted and indexed. No information is known about the processes such as the paper-title merging, the author-name disambiguation or the citation extraction. Therefore, the data set is more or less treated as a black box which makes it difficult to reason about the results displayed in this section.

In Figure 5.1 the number of papers published in a year are depicted for the different domains. A steady increase in the number of papers for each domain can be observed, especially for Chemistry and Biology.

The graphs in Figure 5.1 show that the data is relatively comprehensive up to 2009 after which a sharp decline in the number of publications can be observed. This seems to indicate that more recent papers have not been indexed from all data sources. The MAS dataset contains papers until 2013 but a lot of recent publications are not associated with venues and therefore are not included in this analysis.

Curiously, an abnormal jump in the number of papers can be observed from 1994 to 1995 for most domains. The domains that exhibit this jump the least are Agricultural Science, Arts & Humanities, Economics & Business, and Social Sciences. The reason for this anomaly cannot be explained easily and seems to come from an internal indexing error of the MAS data. This anomaly is exhibited by all papers, independent of which publishing source they were indexed from and can be observed in most figures throughout this section.

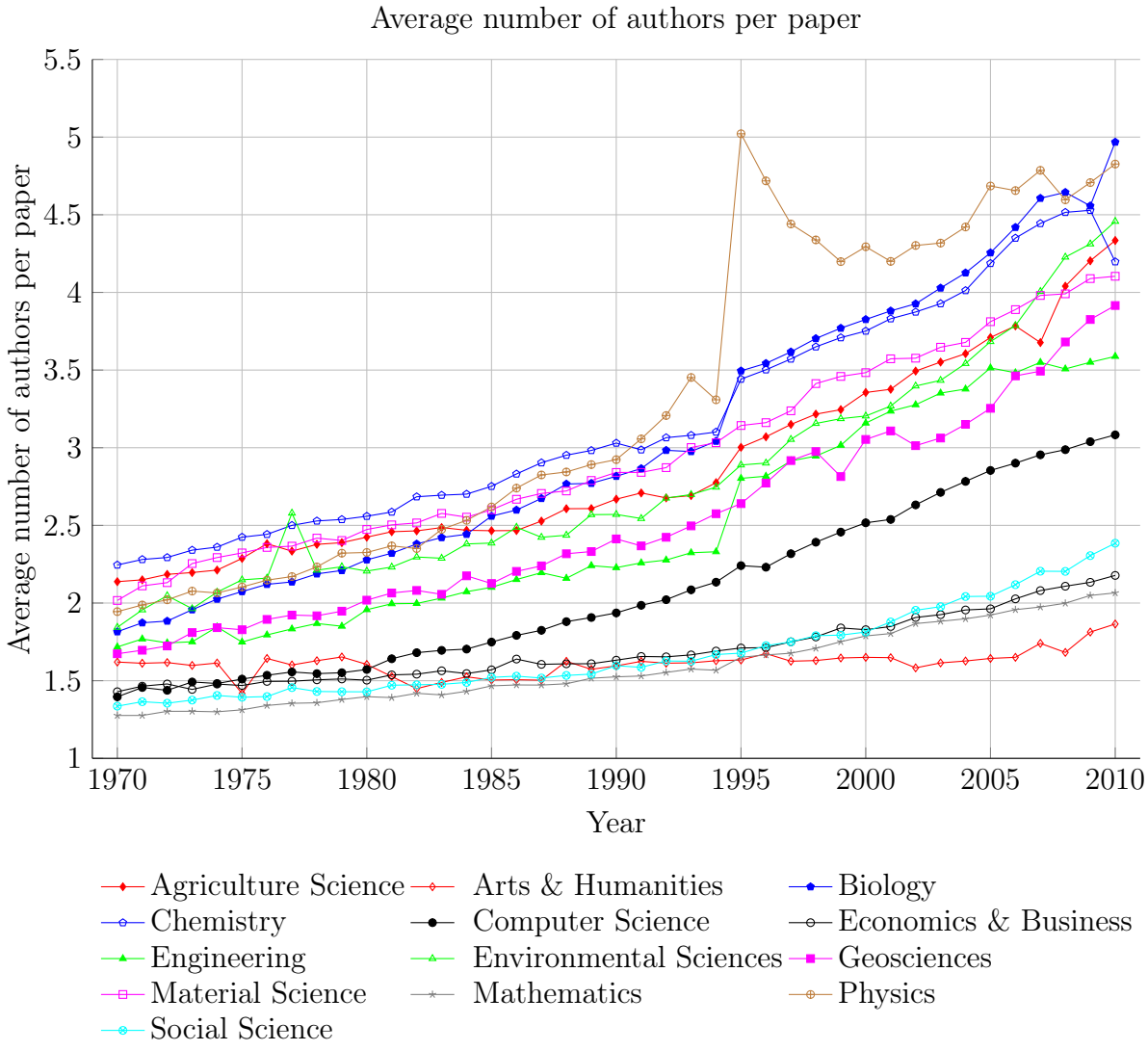


Figure 5.3: The change in the average number of authors per paper over time.

A similar trend as discussed above can be observed in Figure 5.2 where the number

of new authors that publish their first publication are plotted over time. Again, a sharp decline occurs after 2009 and the anomaly can be observed in the data from 1994 to 1995 where a sudden increase in the number of new authors occurs.

Figure 5.3 shows the change of the average number of authors per paper over time. For all domains the average number of authors per paper increases with time. The domains that have the smallest number of authors per paper are Arts & Humanities, Mathematics, Economics & Business, and Social Science. Papers published in these domains have an average of 1.42 authors in 1970 and 2.21 authors in 2010 which is an increase of 55.56%.

All other disciplines exhibit a much steeper increase from 1.87 in 1970 to 4.16 in 2010 authors per paper, which is an increase of 123.23%. The smallest and largest increases in the number of authors per paper in the 40 years is exhibited in the Arts & Humanities (14.81%) and Environmental Science (142.39%), respectively.

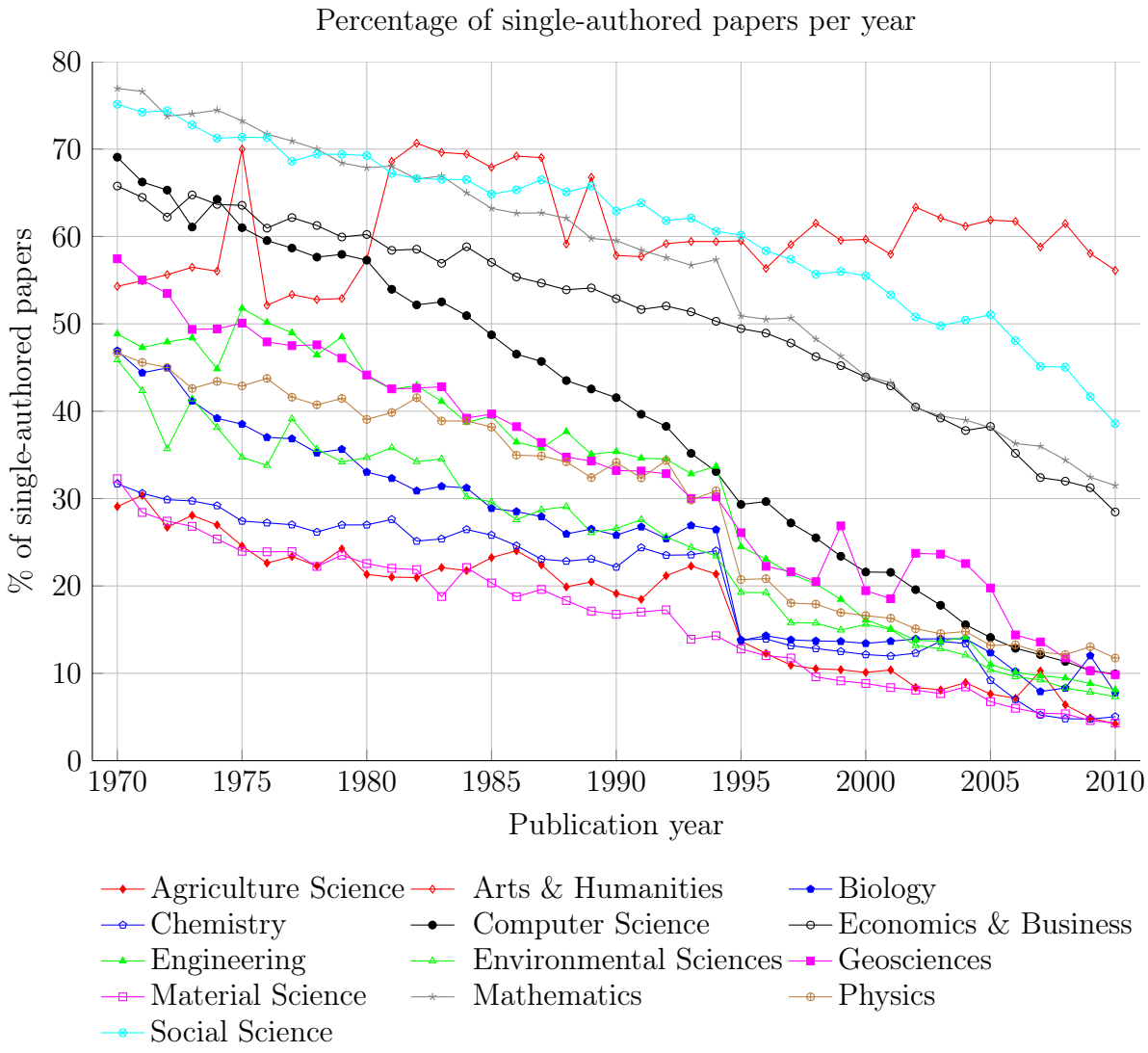


Figure 5.4: The % of single-authored papers over time.

Figure 5.4 shows a complementary graph to the previously discussed figure. Instead of displaying the average number of authors per paper, Figure 5.4 shows the percentage of single-authored papers over time. As expected, the fraction of single-authored papers decreases steadily for most domains. In 1950 the percentage of single-authored papers

over all disciplines is 65.11% while in 2010 it decreases to 17.15%. One can see that Computer Science has the steepest decrease in single-authored papers from 90.82% to 9.92%.

The only discipline in which the percentage of single-authored papers increases is Arts & Humanities with an increase of 10.32% when compared over the 60 year time span. It should also be noted that the data anomaly is also exhibited in this Figure where a jump in the percentage of single-authored papers can be seen from 1994 to 1995.

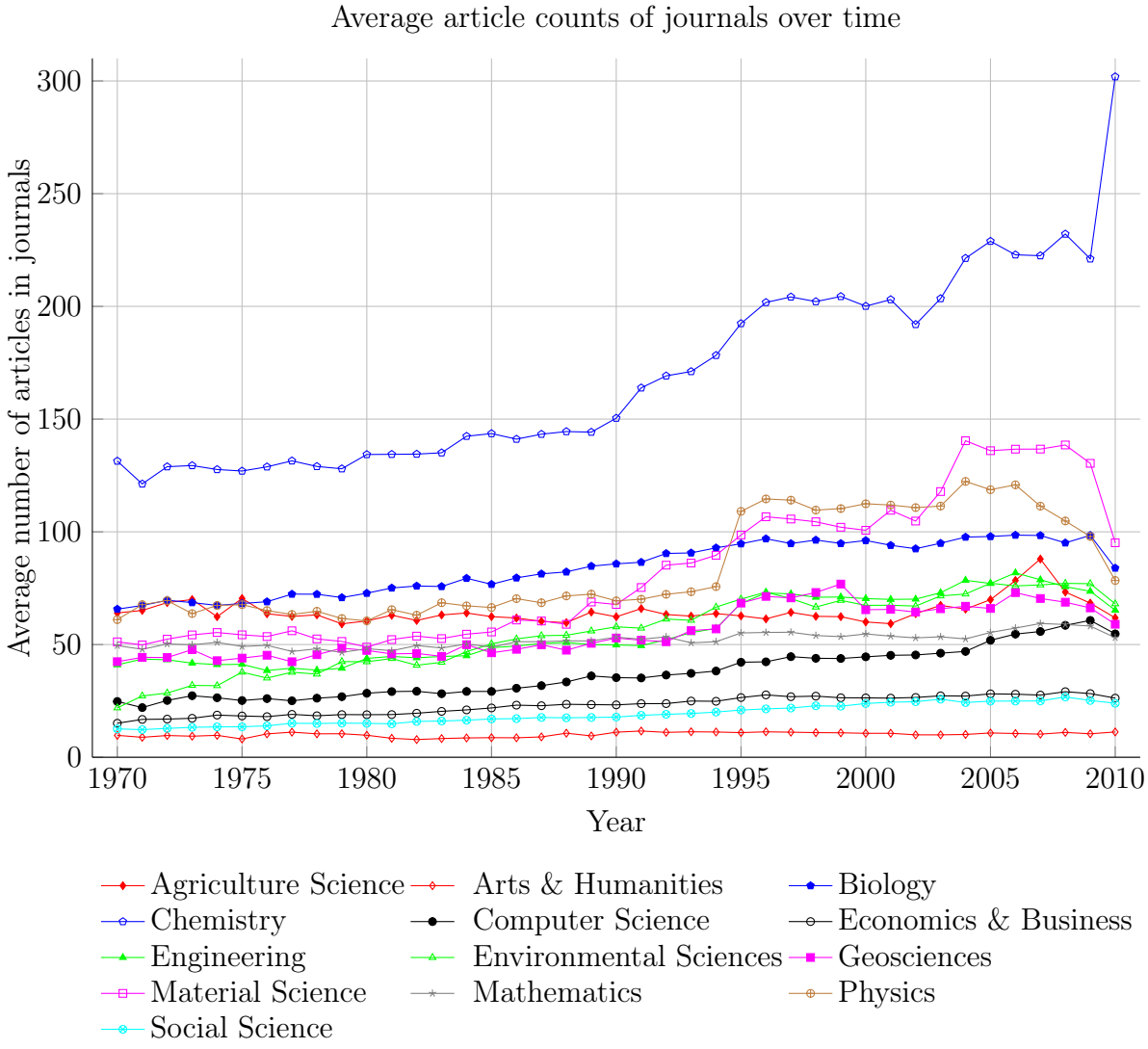


Figure 5.5: The average number of articles published in journals over time.

Figure 5.5 shows the average number of articles that are published in journals over the years. No distinction is made between journals that publish weekly, monthly or yearly. An increase in the number of articles published by journals can therefore be attributed to three factors. Firstly, the time between journal editions decreases. Secondly, the number of articles per edition increases, or thirdly, more papers are indexed per journal by MAS in later years. Unfortunately, volume and edition information is not available for the MAS data since publication dates are not granular enough. In other words, the publication dates of papers are years and not months or days.

By taking the average number of articles published per journals in the years 1950 to 1954 and comparing it to the average number of articles published per journal in the years 2006 to 2010, the increase in the average number of articles published per journal over time can be computed. The domains with the smallest increases are Physics (78.53%), Agriculture Science (80.68%) and Arts & Humanities (84.74%). Similarly, the domains with the largest increases are Social Science (269.71%), Geosciences (291.40%) and Environmental Sciences (325.33%).

Figure 5.6 shows the average number of citations that papers receive plotted against the publication years. It is reasonable to argue that the average citation rate should be constant throughout the years since the number of citable papers grows linearly to the number of citing papers unless the reference lists of papers grow larger over the years. For reference, the change in the reference list sizes over time are plotted in Figure 5.7.

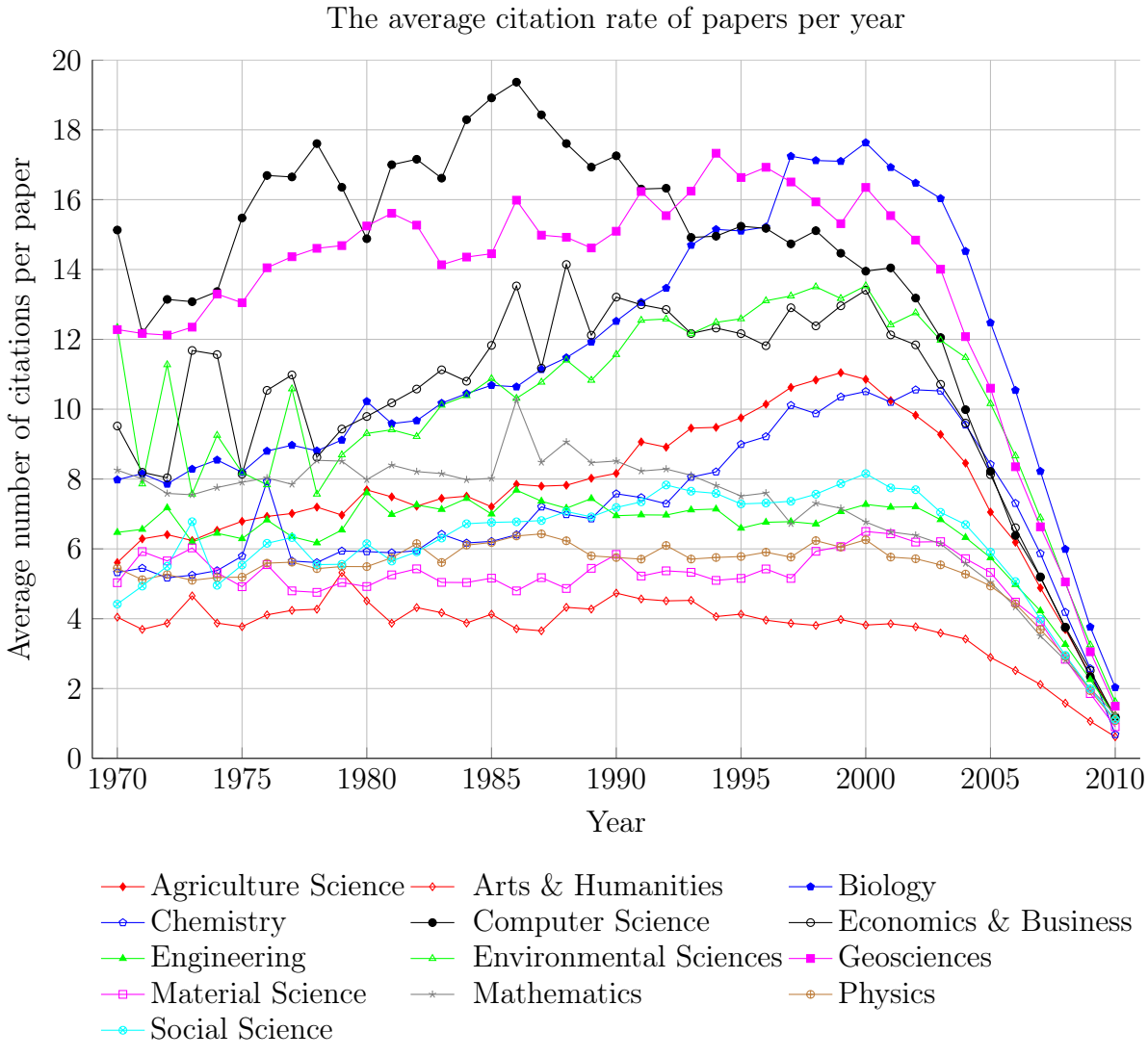


Figure 5.6: The average citation counts of papers over time.

One can see that the citation rate of papers stays relatively stable for most domains between 1970 and 2000 with a slight upward trend. After 2000 a steep decline in the average number of citations per papers can be observed. This decline can be explained

by the decreasing number of papers in later years that are potential sources of citations and the fact that papers are not indexed from 2013 onwards.

Figure 5.7 shows the average number of references that papers have in their reference lists since 1970. In each domain the number of papers that are referenced has steadily increased from 2.19 in 1970 to 18.70 in 2010. Note that a sudden increase in the reference lists of papers can be observed from 1994 to 1995 for all domains.

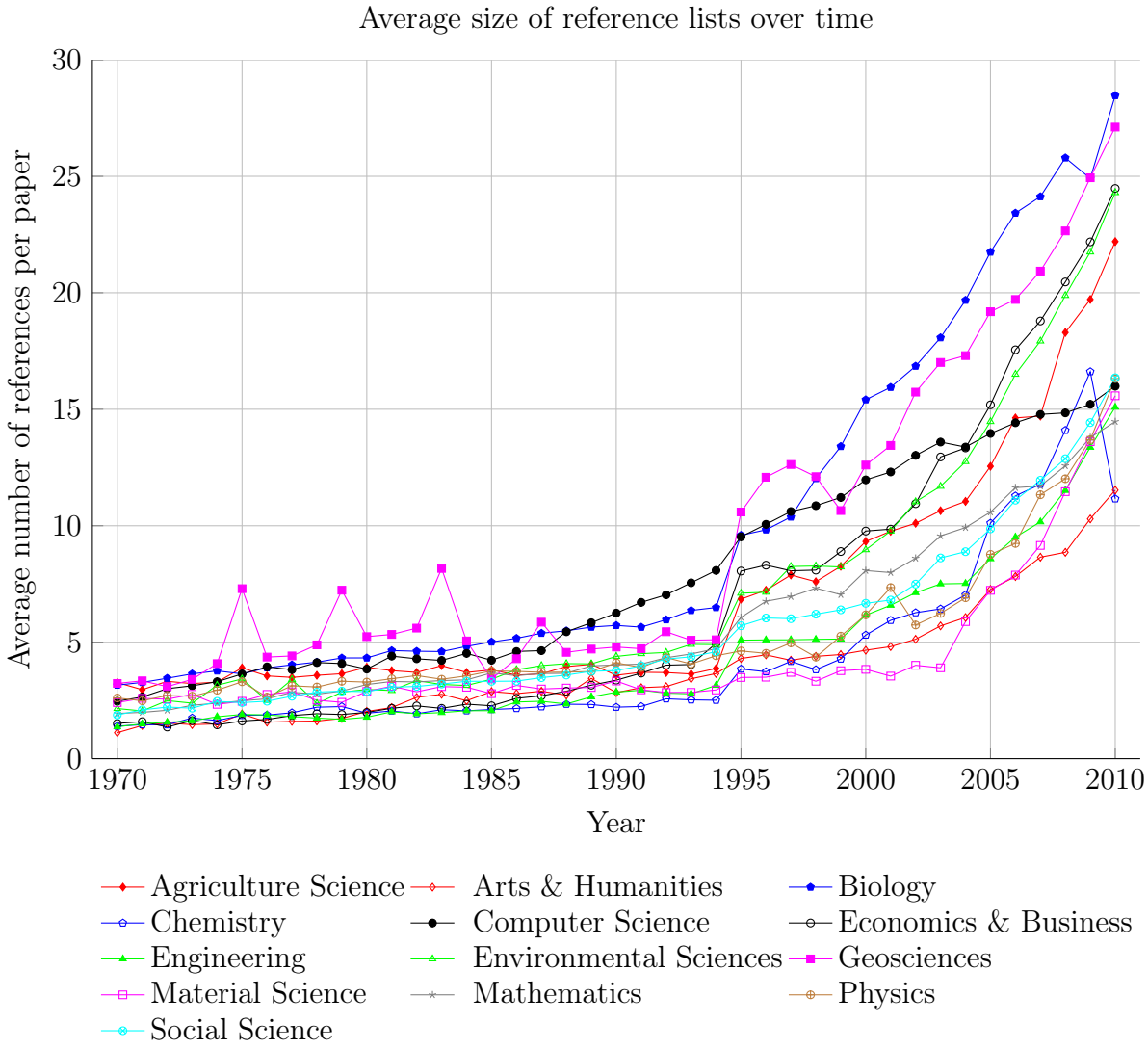


Figure 5.7: The change of the average size of reference lists over time.

Considering the average number of papers in reference lists between the years 1970 and 1974 and comparing it to the average number of referenced papers in the years 2006 to 2010, the average increase in the reference list sizes is computed. According to this comparison, Environmental Science, Chemistry and Economics & Business are the domains with the largest increases. Similarly, the domains with the smallest increases are Material Science, Physics and Computer Science.

It should be noted that the small number of papers in the reference lists of papers published a longer time ago could probably be caused by the lack of indexed papers and therefore a large number of references are not counted.

Figure 5.8 shows the average number of citations that papers receive since their publication. It can be observed that for most domains a peak citation rate for papers is reached after only a couple of years since publication. In general this peak is reached three to six years since a paper's publication, after which a gradual decline in the citation rate is seen.

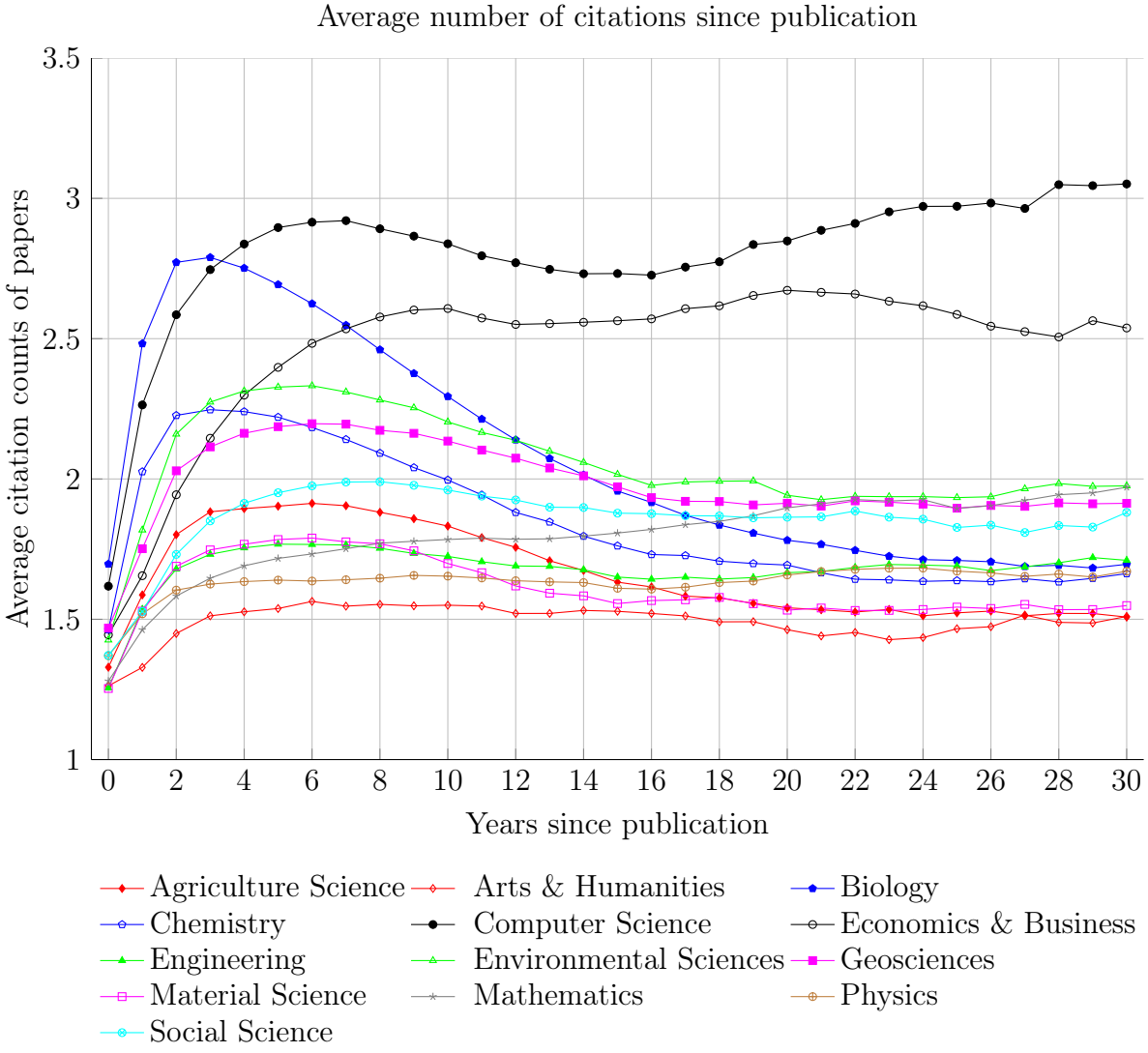


Figure 5.8: The average number of citations per paper since publication.

The only domain where this general trend cannot be observed is Mathematics, where papers receive more citations a year the older they are. This seems to indicate that the life-time of Mathematics papers is longer compared to other fields where results seem to be obsolete more quickly and therefore are not referenced by newer papers anymore.

Both Physics and Arts & Humanities show an initial increase in the number of citations and after about 4 years since publication, papers seems to obtain a stable citation rate of 1.64 and 1.52 on average, respectively.

The domains Economics & Business and Computer Science exhibit a slightly different trend in this figure. Both domains reach a peak in their citation rates after 7 and 10 years, respectively, after which their citation rates decrease. However, it appears that after 12 and 16 years their citation rates increase again. This second increase in the citation rates is not exhibited by any other domain. The reasons for this different behaviour are unclear.

One possible explanation for the Computer Science domain could be the discrepancy between theoretical and practical advances. In other words, theoretical research goes unrecognized until hardware requirements are met to implement the theory. However, further analysis is required to find a definitive answer to this citation behaviour.

Considering only the first 15 years after the initial publication of the papers, the citation peaks for the various domains are reached after different number of years. The amount of time it takes for citation rates to peak are summarised in Table 5.5.

Table 5.5: The number of years it takes for the citation rates of papers to peak in the different domains. Only the first 15 years after the initial publication are considered.

Domain	**Peak Year**	**Citation Peak**
Mathematics	15	1.81
Economics & Business	10	2.61
Physics	9	1.66
Social Science	8	1.99
Computer Science	7	2.92
Agriculture Science	6	1.91
Arts & Humanities	6	1.56
Environmental Science	6	2.33
Geosciences	6	2.20
Material Science	6	1.79
Engineering	5	1.77
Biology	3	2.79
Chemistry	3	2.25

If Mathematics papers are ignored, since their citation rate increases the older papers get, then papers from Economics & Business take the longest to reach their citation peak, namely 10 years, and receive 2.61 citation on average. The domains in which papers reach their citation peaks the fastest are Biology and Chemistry, namely 3 years.

Figure 5.9 shows the number of authors that publish journal articles x years since their first publication. As expected the number of authors that continue publishing journal articles decreases over time. This is to be expected since only a few authors continue publishing after 20 year careers.

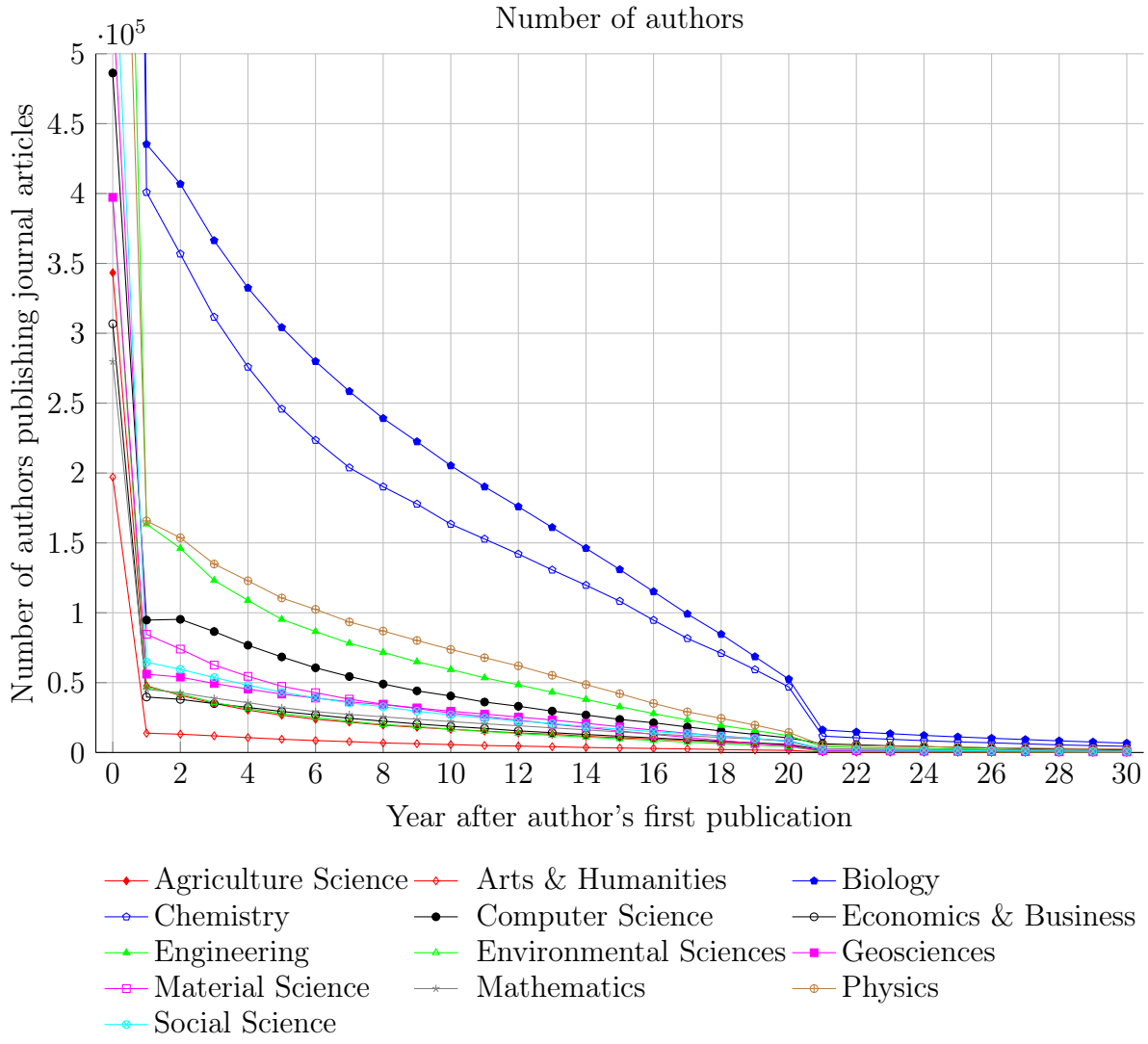


Figure 5.9: Number of authors publishing journal articles since their first publication.

It should be noted that the anomaly mentioned before can be observed again after 20 years. In Figure 5.9 a sharp decrease in the number of authors that publish x years after their first publication can be observed.

Figure 5.10 plots the average number of journal articles that are published by authors since their first publication. It appears that, on average, researchers publish 1.14 articles in the first year of their academic careers. In their second year this value increases to 1.35. After that, a steady increase for most domains can be observed up to 16 to 18 years into a researcher's career. The anomaly that perseveres throughout the MAS data can also be observed in this figure at year 20 after an author's first publication.

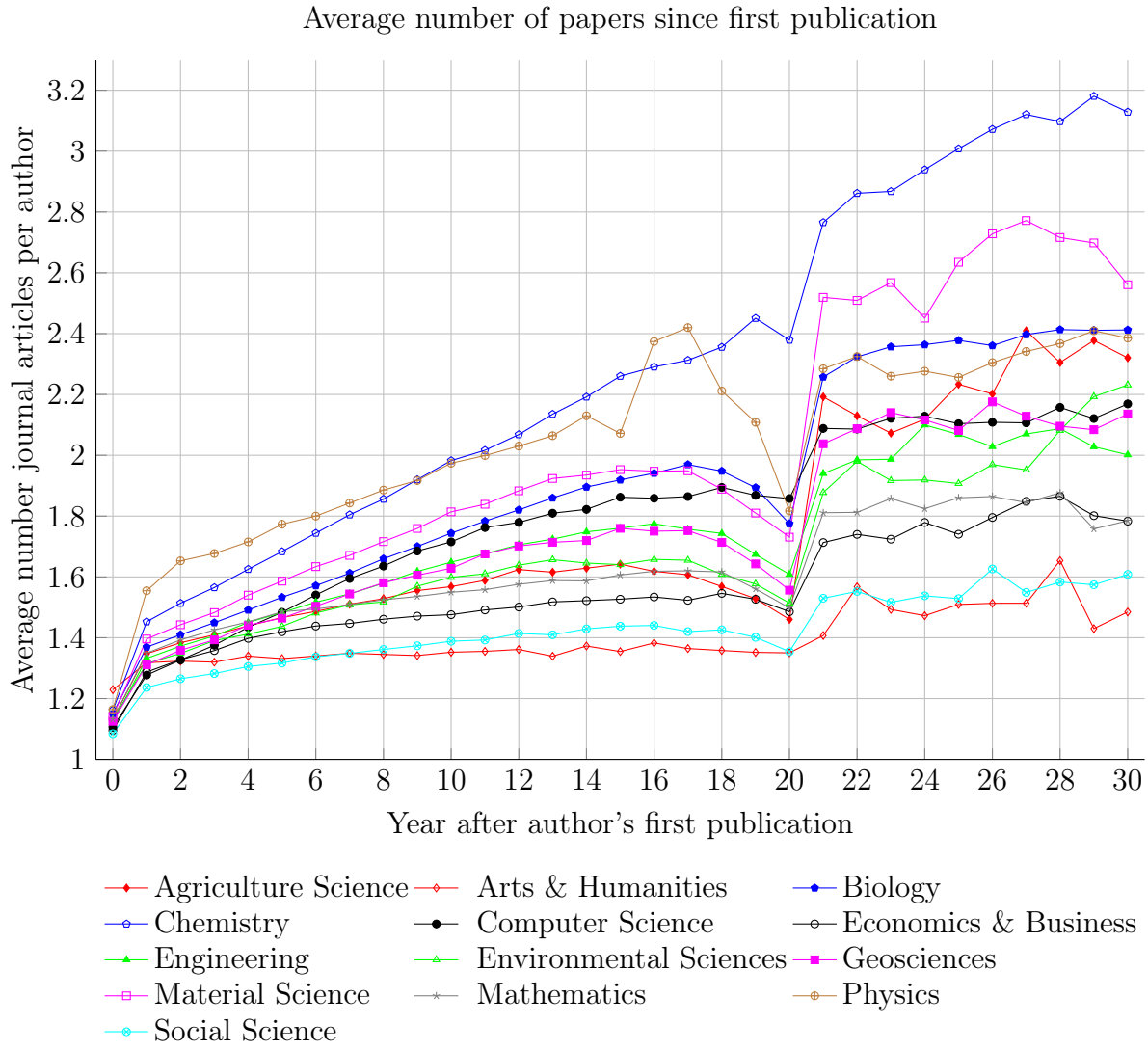


Figure 5.10: Average number of journal articles published by authors since their first publication.

When computing the average values for the years 1 to 3 and 16 to 18 and comparing the increases for the different domains, it is found that authors in Arts & Humanities, Social Science and Agriculture Science have the smallest increase in journal article outputs, with an average increase of 3.65%, 13.28% and 15.78%, respectively. Alternatively, the steepest increases in the publication output is observed by Computer Science, Physics and Chemistry authors, with an average increase of 41.15%, 43.42% and 53.57%, respectively.

Figure 5.11 shows the average ratio of journal to conference papers published by authors since their first publication. If the average publication ratio is 0.5 for year x, it implies that the average number of journal articles and conference articles published by authors x years since their first publication is exactly the same. Therefore, if the value in the graph lies above 0.5, it indicates that, on average, authors publish more journal articles in that stage of their careers. Alternatively, if the value is below 0.5, it means that more conference articles are published by the authors.

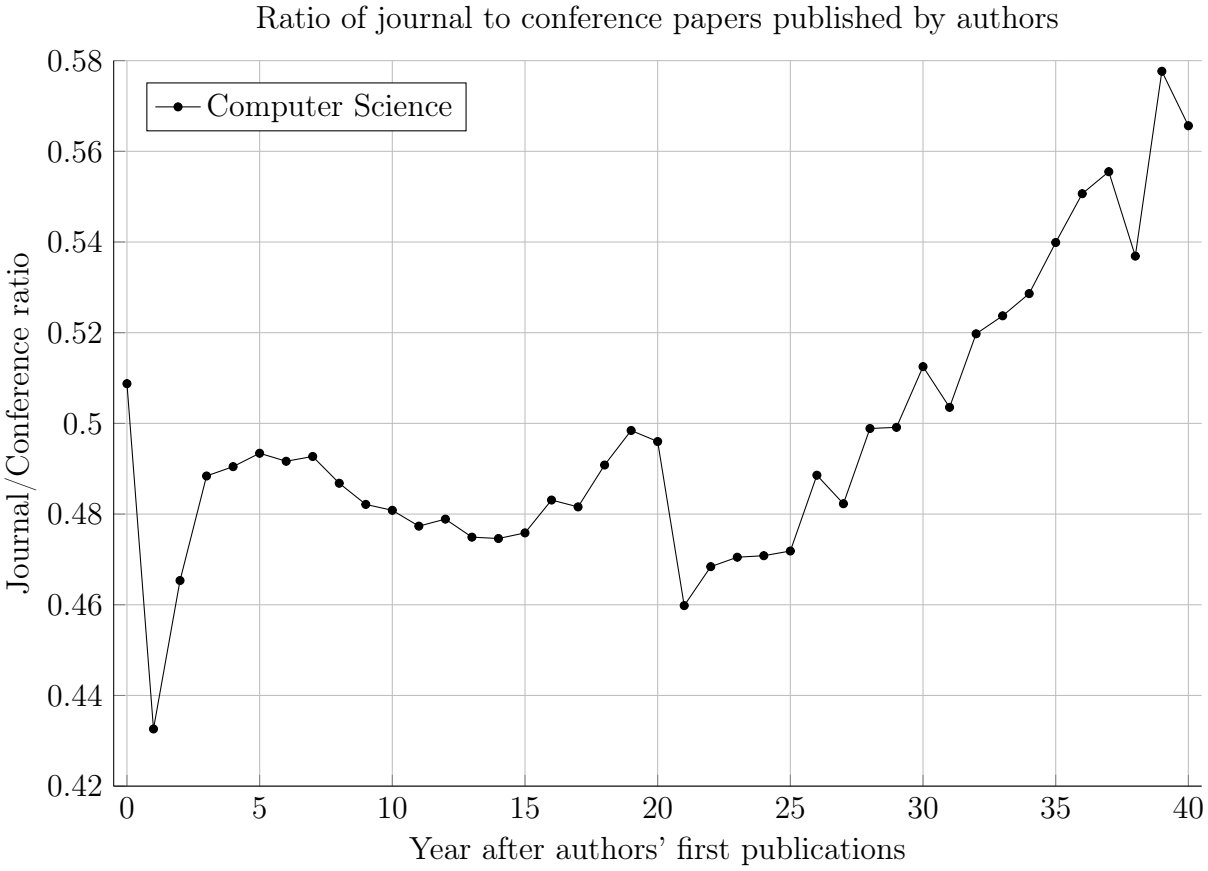


Figure 5.11: Ratio of journal to conference papers published by authors since their first publication.

This trend is only plotted for the Computer Science domain because it is the only domain in which conference articles constitute a large portion of the research output. Engineering and Computer Science are the only domains that contain more than just a few conferences. However, the number of conference articles is dwarfed by the number of journal articles in Engineering where researchers publish around 97% of their articles in journals. The total number of conferences and journals per domain are listed in Table A.1 in Section A.1.

Again, an anomaly in the data can be observed 20 years since the first publication of authors. Therefore, one has to consider the plot in Figure 5.11 in two parts. The values for the years 0 to 19 since an authors first publication have to be considered independently from the values for the years 21 to 40.

It appears that Computer Scientists' first publications are rather journal articles than conference articles. However, one can see that Computer Scientists publish more confer-

ence articles than journal articles in the first years of their careers. Only later, after 30 years, do Computer Scientists publish more journal than conference articles on average.

Figure 5.12 shows the average age of papers that are cited in a year. In other words, the average time between the publication years of the papers that are contained in reference lists of papers and the referencing papers are given. It appears that the average age of the references stay roughly the same over the years, with a slight upward trend.

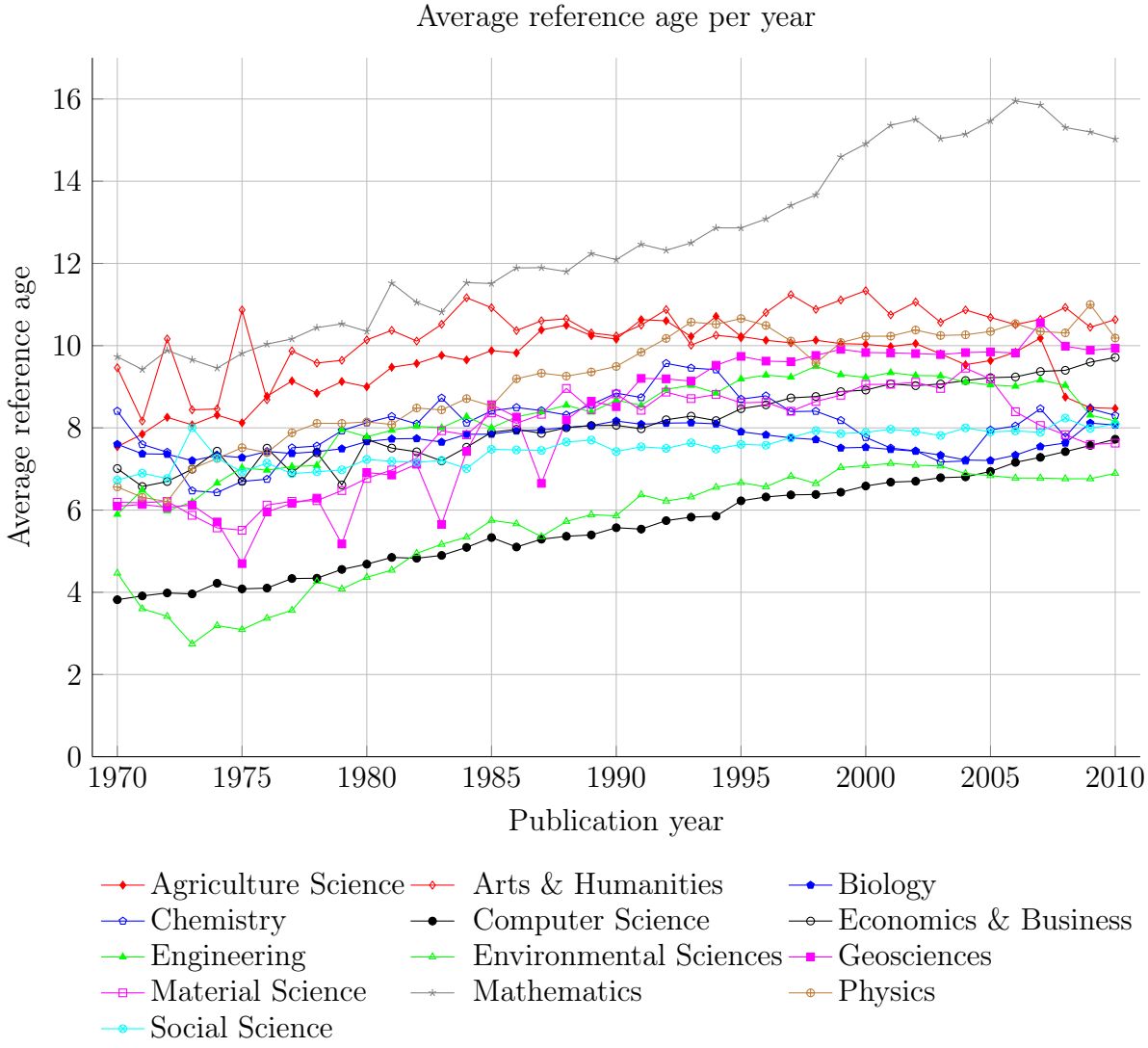


Figure 5.12: The average age of the papers that are referenced in a year over time.

The domains where the age of references increases the most are Physics, Geosciences and Mathematics with an increase of 3.81, 4.01 and 5.84 years, respectively. Similarly, the domains where the age of references stay the most constant are Biology, Social Sciences and Chemistry with an increase of 0.38, 0.90 and 0.95 years, respectively.

5.5 Chapter Summary

This chapter described the data sets that are used to construct citation networks and how the data is cleaned up in order to obtain coherent data. It was found that the DBLP data

set has a very low coverage in the citation data with a high precision. A citation network constructed from the DBLP data set should therefore only be used for comparison reasons.

In addition, the evaluation data sets that were collected were discussed in this chapter and how these data sets are used to evaluate the performance of the ranking algorithms and publication venues in predicting high-impact papers.

Lastly, properties of the MAS data were given and some publication trends that differ between different academic domains were identified. It was found that there exists at least one data anomaly in the MAS data. However, the source for this anomaly could not be identified.

Chapter 6

Comparing Ranking Algorithms

In this chapter the outputs of the algorithms, which are lists of rankings, are compared empirically to identify the algorithms' ranking properties, strengths and weaknesses.

Chronologically this chapter is divided into three parts in which the ranking algorithms that rank different entities are analysed. For instance, Section 6.1 covers the paper ranking algorithms. The algorithms for venues and authors are analysed and compared in Sections 6.2 and 6.3, respectively.

6.1 Comparing Paper Ranking Algorithms

The algorithms that rank individual papers are compared using the MAS Computer Science citation network as input. This is not a trivial task because of the size of the data and the varying purposes of the algorithms. The approaches used in this section are intended to answer the following questions:

- Are there significant differences in the convergence speeds of the algorithms?
- How much do the rankings of the papers produced by the algorithms differ? Are there similarities or disparities between the algorithms when looking at their output?
- What are the characteristics of the top ranked papers according to each algorithm?
- Are there properties of the algorithms that can be identified by looking at papers that are outliers when the rankings of the algorithms are compared using scatter lots?
- How do the algorithms distribute the scores of papers over the publication years? In other words, do the publication years of papers play an important role when the algorithms are used on bibliographic citation networks?

For all results presented in this section the parameters of the algorithms were set to their default values as indicated by the authors introducing the methods, unless specifically stated otherwise. Therefore, the damping factor α of PageRank was set to 0.85 which is also used for YetRank, SceasRank and NewRank. Similarly, the time decay parameter used by YetRank and NewRank was set to $\tau = 4.0$. The target and census window sizes, which are used by the Impact Factor method in YetRank, were set to 5 and 1 years, respectively. The two additional parameters of the SceasRank method, a and b

were set to e and 1. Lastly, the precision threshold was kept the same for all algorithms and set to $\delta = 10^{-6}$.

For the experiments in the following sections, the MAS CS subset was used to construct the citation network of 2 394 976 vertices and 12 907 440 edges. Where it seemed appropriate to use an additional set of data, the DBLP citation network was used. This was done to determine whether results depend on the characteristics of the underlying data or not.

6.1.1 Convergence Rates of the Algorithms

Academic citation networks are much smaller compared to hyperlink graphs of the world wide web but can still contain millions of vertices and edges. Therefore, the computation times and convergence speeds of the algorithms are important.

In Figure 6.1 the convergence speeds of the ranking algorithms with time complexities of $O(n)$ per iteration are given. The CiteRank algorithm is not included since its cost is close to $O(n^3)$ and would dwarf the other results. The x-axis is the number of iterations required to achieve a precision of δ or higher, given by $\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|_1$ which is the grid distance between the result vectors of successive iterations.

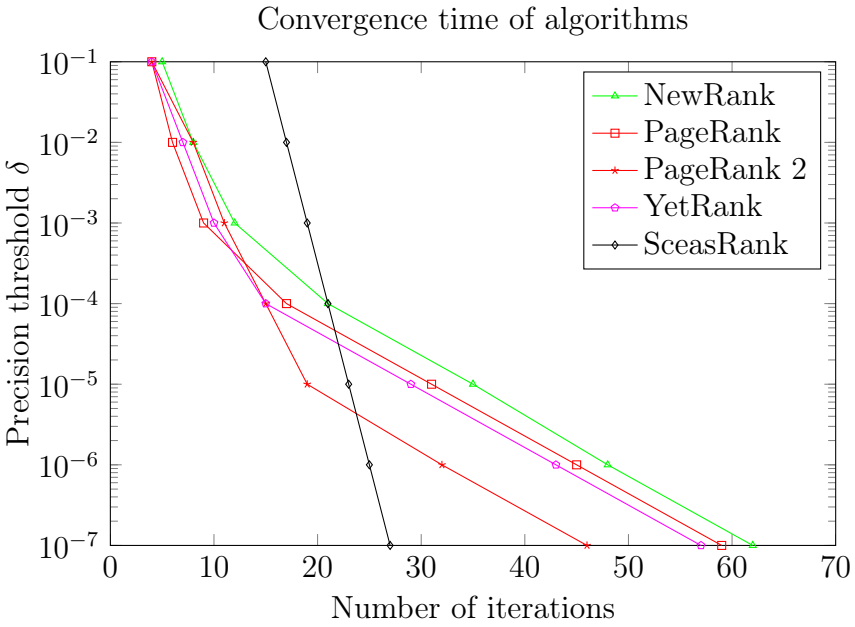


Figure 6.1: Convergence speeds of the ranking algorithms, initialised with the default parameters, on the MAS CS citation network. For comparison reasons PageRank 2 shows the convergence rate of PageRank when no additional edges are added to dangling vertices in the citation network.

The precision threshold criteria should be defined separately for each algorithm and depends on the expected magnitude of the result vector and the underlying citation network. For example, PageRank-like algorithms that add N edges to dangling vertices and use a damping value α in the range of $(0, 1)$ converge to a result vector with magnitude 1. On the other hand, the magnitude of the result vectors of SceasRank and CiteRank depend on the size of the network that is used in their computations. Moreover, as mentioned in Section 2.5.1 and shown in Figure 7.3, the computation time of PageRank-like

algorithms also depends on the damping factor α. Therefore, the same value of $\alpha = 0.85$ is used for all algorithms in this comparison. In order to compare the convergence speeds of the algorithms one has to look at the slopes of the lines in Figure 6.1. The smaller the slope of a line, the faster the corresponding algorithm converges. One can clearly see that SceasRank behaves differently from the other algorithms and converges much faster.

The reason for the large number of iterations used by SceasRank with a relatively small precision threshold is that the sum of the result vector does not have an upper bound of 1 and initially contains large variances. All other algorithms have approximately the same convergence speeds.

It should be noted that other aspects influence the total computation times of the algorithms. YetRank, for example, has an expensive initial overhead computation since the Impact Factors for all venues and each year under consideration have to be computed.

6.1.2 Correlation between Paper Ranking Algorithms

In order to quantify the similarity between the rankings produced by the different algorithms, three different correlation measures are used. The Pearson correlation coefficient r measures the linear dependence between two variables X and Y and returns correlation values ranging from -1 to 1. A correlation of 1 implies a perfect linear correlation between X and Y, where all data points lie on a line for which Y increases as X increases. The Pearson correlation coefficient for a sample is represented by the letter r and is formally defined as:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{6.1.1}$$

where n is the sample size and $\overline{X}$, $\overline{Y}$ are the sample means of each variable. The Pearson value r is known not to be robust for data that contains outliers and therefore can be misleading. This is a problem with heavy-tailed data such as the in-degree distribution exhibited in citation networks. More importantly, the Pearson correlation depends on the actual score values of the results which is not important in the rankings of the elements.

The Spearman and Kendall rank correlations overcome these problems since they compute correlation values between relative ranks instead of absolute values. Let x_i and y_i be the ranks of the elements in the ordered variables X and Y, respectively. The Spearman rank correlation ρ is used to describe the monotonic relatedness between two variables by calculating the Pearson correlation coefficient between the ranks x and y. It is defined as

$$\rho = \frac{\sum_i(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i(x_i - \overline{x})^2 \sum_i(y_i - \overline{y})^2}} \tag{6.1.2}$$

On the other hand, the Kendall correlation is computed over each pair of two lists of ranked elements. It counts the difference between the number of concordant pairs and the number of discordant pairs. A pair is concordant iff $x_i > x_j$ and $y_i > y_j$ (or $x_i < x_j$ and $y_i < y_j$). Contrarily, a pair is discordant iff $x_i > x_j$ and $y_i < y_j$ (or $x_i < x_j$ and $y_i > y_j$). Lastly, a tied pair occurs if $x_i = x_j$ or $y_i = y_j$. Formally, Kendall's Tau-b (τ) value is defined as

$$\tau = \frac{N_c - N_d}{\sqrt{(N_c + N_d + N_t)(N_c + N_d + N_u)}} \tag{6.1.3}$$

where N_c and N_d are the number of concordant and discordant pairs. N_t and N_u are the number of ties in X and Y, respectively. The Kendall τ value is typically used to quantify

the rank stability and the rank similarity between two variables with 1 indicating perfect agreement between the two rankings and -1 reflects that one ranking is the reversal of the other.

Number of Common Elements in Top Rankings

Given the ranking outputs of all algorithms on the MAS CS citation network, the top 50 papers are extracted. The reason for only considering the top papers in this comparison is twofold. Firstly, the top results are the most important for ranked elements in any type of information retrieval application. Secondly, the ranking algorithms considered in this paper introduce a lot of noise by elements that are not highly ranked. Nonetheless, the correlations between the algorithms when considering all papers in MAS CS citation network are given later in this section for comparison.

The number of papers that are common in the top 50 rankings for each pair of algorithms are displayed in Table 6.1. In addition, the Spearman and Kendall rank correlations coefficients are given since they measure the similarity of the rankings more accurately.

Table 6.1: Number of common papers in the top 50 rankings of each algorithm and the associated rank correlation coefficients (ρ, τ). For each algorithm the average publication year, the average number of citations and the year range in which the top 50 ranked papers were published are also depicted. PageRank and SceasRank have the highest correlation according to the Spearman ($\rho = 0.76$) and Kendall ($\tau = 0.60$) rank correlation coefficients. PageRank and SceasRank also have the highest number of common papers (38) in the top 50 ranked papers. The lowest correlation is found between YetRank and NewRank with $\rho = 0.09$ and $\tau = 0.07$. These two algorithms also have the smallest number of common papers in the top 50 ranked papers.

	CountRank	**PageRank**	**NewRank**	**YetRank**	**SceasRank**
CountRank	–	29	20	27	26
PageRank	(0.67, 0.47)	–	31	19	38
NewRank	(0.56, 0.38)	(0.41, 0.27)	–	6	28
YetRank	(0.53, 0.37)	(0.51, 0.36)	(0.09, 0.07)	–	11
SceasRank	(0.52, 0.32)	(0.76, 0.60)	(0.49, 0.35)	(0.18, 0.09)	–
Avg. Year	1990.58	1989.02	1996.26	1988.96	1990.36
Avg. Cites	3507.88	2923.28	2141.12	2518.02	2854.02
Year Range	[1963, 2010]	[1960, 2010]	[1963, 2010]	[1970, 2001]	[1960, 2010]

Throughout this section, table cells are highlighted or framed to point out respectively high or low similarity between two algorithms. In Table 6.1, for example, PageRank and SceasRank have the highest number of common elements, namely 38, in the top 50 ranked papers. They also have the highest Spearman and Kendall correlation coefficients of 0.76 and 0.60, respectively. This is to be expected, since SceasRank is the most similar to PageRank since it does not incorporate publication dates and venue impact factors in its computation and, in addition, only adds weights from papers that have a score of zero.

From the rank correlations in this table one can see that the top rankings of PageRank have the highest correlation with the top rankings produced by CountRank. Empirically this observation makes sense due to the fact that PageRank is the algorithm that models the citation network the most closely to citation counts and uses the least additional information, such as publication years or venue impact factors, for computing ranking scores.

This is also supported by the average number of citations that the top 50 ranked papers received. The top papers, according to PageRank, have an average of 2923.28 citations which is the closest to CountRank. A slightly smaller number is given by SceasRank with 2854.02 citations per top paper which further shows its similarity to PageRank.

A high number of common papers (31) is also produced by PageRank and NewRank. This high overlap indicates that a high number of citations to a paper (2141.12 on average) outweighs the impact that the publication dates have on the scores of the top papers. Nonetheless, the average publication year of the top 50 papers according to NewRank is 1996.26 which is still considerably later than for the other algorithms which average around 1990.

The two algorithms that are the least similar in ranking the top 50 papers are NewRank and YetRank. The low correlation of only 6 common elements is surprising since the two algorithms are very similar in nature except that YetRank includes the Impact Factors of venues in its computation. YetRank puts the average publication year at 1988.96 which is close to CountRank (1990.58) and PageRank (1989.02) but not NewRank which puts the average publication year of the top 50 papers at 1996.26. This observation, however, does not explain the low correlation between the two algorithms. The reasons for a low number of common papers in the top rankings become clearer in Section 6.1.3.

Curiously, while the other algorithms' top 50 lists include papers up to 2010, YetRank only lists papers published until 2001. Although YetRank includes the publication dates of papers in its computation, it seems that the impact factors of the venues outweigh the impact that the publication dates have on the score of the top papers.

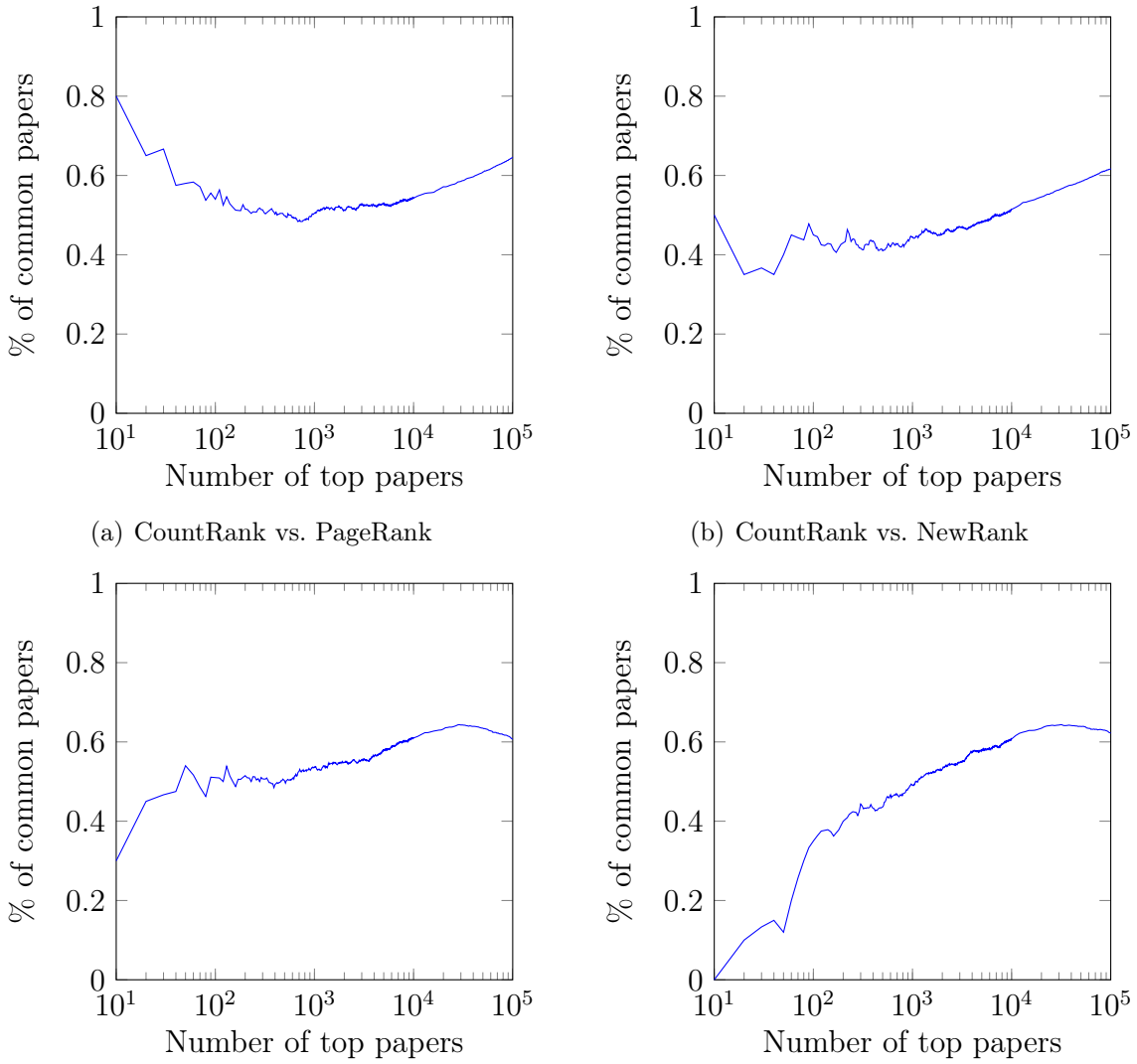
Stability of Rankings

It can be argued that a comparison between the top 50 rankings of each algorithm is skewed by outliers (high number of citations) and does not give a true indication of the similarity between the different rankings. A more comprehensive picture can be obtained by counting the number of common elements in the top rankings with changing sample sizes. In other words, how does the similarity between two algorithms change when considering a different number of top ranked papers?

For this, let $Top(x)$ be the number of common elements in the top x rankings of two algorithms. The percentage of common elements in the top x rankings is then given by $Top(x)/x$. Some comparisons using the number of common elements in the top rankings with varying sample sizes are given in Figure 6.2, where $Top(x)/x$ is plotted against x.

From the graphs in Figure 6.2 one can see again that the PageRank rankings are the most similar to counting the number of citations to papers. Both PageRank and NewRank show a steady increase in the similarity to CountRank the larger the sample size becomes. This is different for YetRank which shows a decline in similarity after the sample size reaches 30 000 papers.

The percentage of common elements in the rankings of the algorithms tend towards 60% for the varying sample sizes. It is surprising that the correlation is this low even for small sample sizes and cannot be explained easily. It should be noted that noise is introduced for papers that are ranked low and would explain low correlations for large sample sizes. On the other hand, the percentage of common elements has to tend towards 100% when the sample size reaches the size of all papers in the data set.



(a) CountRank vs. PageRank

(b) CountRank vs. NewRank

(c) CountRank vs. YetRank

(d) NewRank vs. YetRank

Figure 6.2: The percentage of common papers in the top rankings of the different algorithms. The number of top papers considered is given on the x-axis.

Correlation between Algorithms

Table 6.2 lists the Spearman and Kendall rank correlation values between the complete rankings of CS papers for each pair of ranking algorithms. As before, the correlation coefficients indicate that PageRank and SceasRank produce the most closely related rankings and that PageRank is the algorithm that is the most similar to CountRank.

On the other hand, when comparing PageRank to NewRank and YetRank using the correlation coefficients they show an opposite picture to using the number of common elements in the top 50 rankings. This time YetRank and NewRank produce rankings that are more similar while the correlation between PageRank and NewRank is very low.

It should be noted that when comparing CountRank to NewRank and YetRank, the τ values are very close together (0.332 and 0.331) while the ρ values differ more (0.474 and 0.468). Spearman's ρ is more sensitive to large discrepancies between rankings than Kendall's τ. Therefore, it seems reasonable to expect that YetRank has more outliers than NewRank when compared to CountRank.

Table 6.2: Rank correlation coefficients (ρ, τ) for the complete rankings of the CS domain for each pair of algorithms. Highlighted cells indicate a high correlation while boxed cells show low correlations between two algorithms.

	PageRank	**NewRank**	**YetRank**	**SceasRank**
CountRank	(0.92, 0.78)	(0.47, 0.33)	(0.47, 0.33)	(0.92, 0.78)
PageRank	–	(0.46, 0.33)	(0.40, 0.28)	(0.99, 0.95)
NewRank	–	–	(0.78, 0.60)	(0.47, 0.34)
YetRank	–	–	–	(0.39, 0.28)

6.1.3 Comparison using Scatter Plots

It is possible to show a direct comparison between the various algorithms by using scatter plots. The scatter plots in Figure 6.3 depict the ranks of papers of various ranking algorithms plotted against the citation counts of papers. In plot 6.3(a), for example, the y-axis indicates the PageRank ranks with low ranks plotted at the top and high ranks at the bottom. Similarly, the x-axis indicates the citation counts of the papers where papers with high citation counts are plotted to the right.

The shapes of the plots are not easy to explain due to the intricacies of the algorithms. More importantly, there are clear outliers which can help our understanding of the algorithms. Therefore, the rest of this section discusses the outliers in detail.

The outliers are colour-coded with different colours depending on where in the plot they lie. The selection of the outliers is somewhat arbitrary but papers with a high rank to citation count ratio are highlighted in red while outliers that have a low ratio are highlighted in green. The data points in black are outliers that are selected by hand in order to obtain more information about these papers.

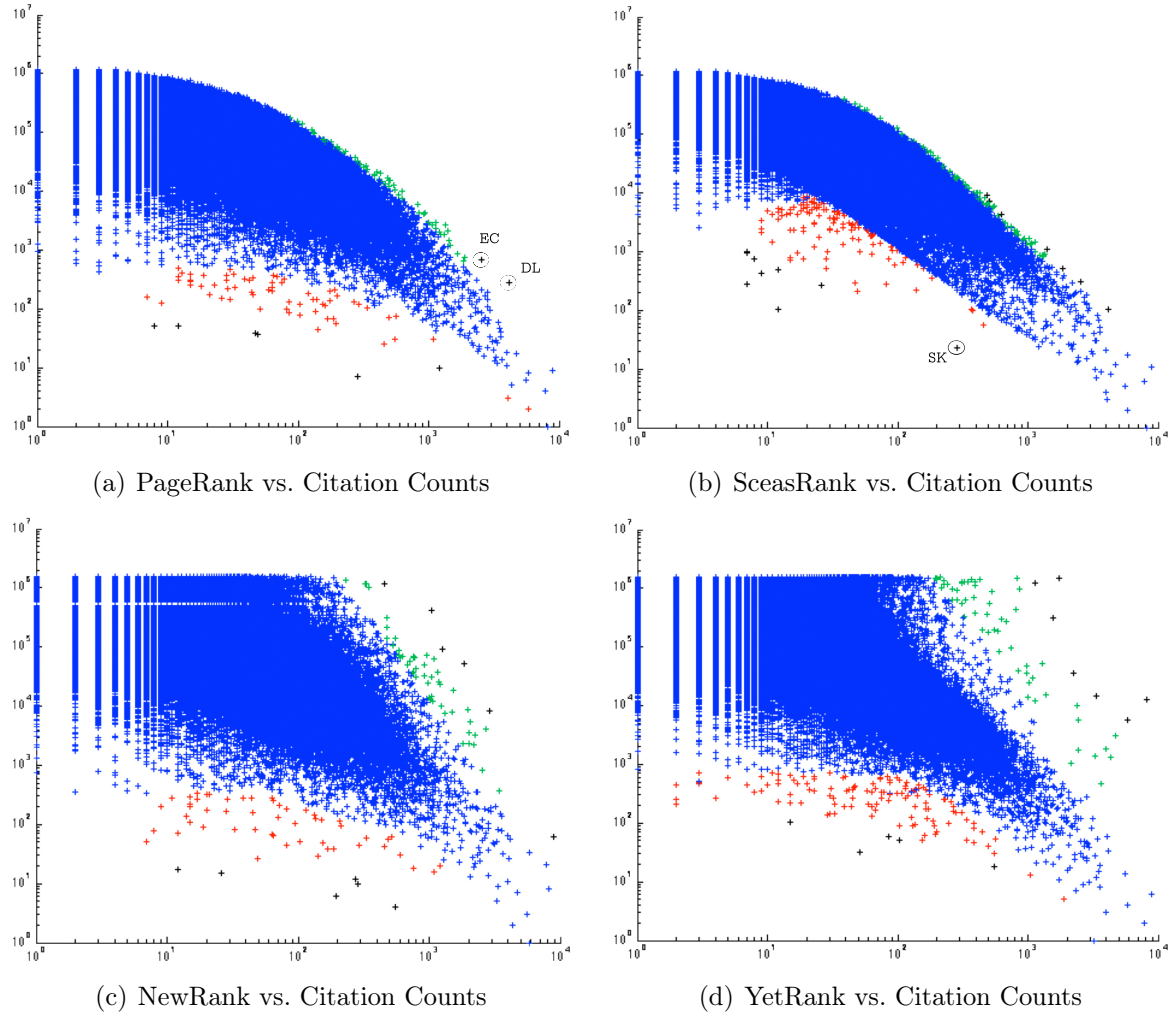
PageRank vs. Citation Counts

Figure 6.3(a) shows the PageRank ranks of papers plotted against their citation counts. The bottom outliers (red) are 72 papers that have relatively high ranks, on average 183, but an average of 280 citations only. The top outliers (green) are 60 papers that obtained a relatively low rank (31 630) but have a lot of citations (705).

The average publication year for the bottom and top outliers is 1985 and 2002 respectively. Therefore, papers with a high PageRank but a relatively low citation count are older papers.

Furthermore, the average Impact Factor for the bottom and top outliers is 0.69 and 3.69, respectively. Both CountRank and PageRank do not take the venue at which papers are published into consideration, yet the citation counts of papers seem to be more aligned with the Impact Factors of the venues at which they are published. In other words, papers have a higher citation count relative to their ranks according to PageRank if they are published at high-impact venues.

The two papers highlighted in black at the top of the graph are Edmund Clarke's (EC) paper "Model Checking", which is the initial paper introducing the method of model checking, and David Lowe's (DL) paper "Distinctive Image Features from Scale-Invariant Keypoints", in which he introduced the scale-invariant feature transform which is one of the most popular algorithms in the detection and description of image features. The citation counts of EC and DL papers are 2567 and 4157, respectively.



(a) PageRank vs. Citation Counts

(b) SceasRank vs. Citation Counts

(c) NewRank vs. Citation Counts

(d) YetRank vs. Citation Counts

Figure 6.3: The ranks of papers for PageRank, SceasRank, NewRank and YetRank plotted against their citation counts. The ranks are computed with the default parameter values for all algorithms. The y-axis indicates the ranks with the first rank at the bottom. The x-axis indicates the citation counts where the papers with the highest citation counts are plotted on the right. The red and green data points indicate outliers with low and high rank to citation count ratios, respectively. The black data points are far outliers that are used to obtain further insight into the ranking properties of the algorithms.

Note that these two papers receive a relatively low rank according to PageRank despite their prominence and fall outside the main body of the scatter plot. A low PageRank rank with many references can only be explained by a large number of references from papers with low ranks. The average scores of the papers citing EC's and DL's papers are 2.73^{-7} and 2.30^{-7} respectively. Comparatively, the average scores of the papers citing the bottom black outliers range between 4.2^{-6} to 5.0^{-5}. In other words, these two papers have a lower than expected rank because the papers that cite them are considered less important by PageRank.

Considering only the 6 black outliers in the bottom of Figure 6.3(a), 5 papers belong to the "Bioinformatics" journal while one belongs to the "Mathematics of Computation" journal. Because of the topics of these two journals, PageRank seems to rank papers higher, compared to CountRank, if they are cited often by non-domain papers. Non-domain papers in the CS citation network are papers that directly cite one or more CS

papers but do not belong to the CS domain themselves. Therefore, they do not have many incoming citations in the network since these citations are truncated.

In order to verify this assumption, the percentage of references from non-domain papers are calculated. For the bottom black outliers the percentage of references from non-domain papers is 65%. The percentages of references from non-domain papers to the papers of EC and DL are 4% and 7%, respectively. Similar values are observed if all bottom (red) and top (green) outliers are considered. The average percentage of references from non-domain papers to the bottom and top outliers is 70.49% and 4.71%, respectively.

Therefore, it seems that PageRank ranks papers higher that are referenced by many non-domain papers. Furthermore, papers that have high ranks but relatively low citation counts are older papers. In Section 6.1.4 more details on how PageRank ranks papers according to their publication dates are given. In addition, Section 7.1 discusses how the scores of papers are distributed differently over the years depending on the value of PageRank's α parameter.

SceasRank vs. Citation Counts

Comparing the scatter plots of PageRank (Figure 6.3(a)) and SceasRank (Figure 6.3(b)), SceasRank seems to have a higher correlation with the citation counts than PageRank even though their Spearman rank correlation coefficient are nearly identical as shown in Table 6.2.

The results of analysing the different outliers are very similar to PageRank. The 83 top outliers (green) have an average publication year of 1998, the venues at which the papers are published have an average Impact Factor of 2.88, and the percentage of references to these papers that originate from non-domain papers is 2.67%.

Considering the 127 bottom outliers (red), the average publication year is 1989, the average Impact Factor of the associated venues is 0.77, and the portion of references from non-domain papers is 64.48%.

Again, most of the bottom outliers are papers published at journals that are not intrinsic to the CS domain. Of the 127 bottom outliers, 33 papers belong to the "Bioinformatics" journal. The second and third most appearing journals are the "Journal of Molecular Graphics" and the "Journal of the ACM", both of which have 5 papers that belong to the bottom outliers.

The furthest outlier at the bottom is Sudhir Kumar's (SK) paper "MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers" published at the "Bioinformatics" journal. It has 284 citations of which 274 citations are from non-domain papers.

It seems that SceasRank behaves similarly to PageRank in ranking papers relatively highly if they have a lot of citations from non-domain papers.

NewRank vs. Citation Counts

The properties of the papers associated with the outliers in Figure 6.3(c) are different to the outliers discussed before.

The 59 bottom outliers, which are papers with a relatively high NewRank rank for their citation count, have an average publication year of 1998 which is much later than PageRank (1985) and SceasRank (1989). In addition, the average Impact Factor of the associated journals is 1.37 compared to PageRank (0.69) and SceasRank (0.77).

When considering the top outliers, papers that have a low rank but relatively high citation counts, the biggest difference to the previous scatter plots is that the papers are

very old with an average publication year of 1975 compared to PageRank (2003) and SceasRank (1999). This was to be expected since NewRank incorporates the publication years of papers into the computation and ranks old papers lower than recently published papers.

Of the 59 bottom outliers, 22 are from the "Bioinformatics" journal. Contrarily, considering the top 56 outliers, the associated papers are predominantly from the ACM. The four most common venues are "Communications of the ACM" (9), "Journal of the ACM" (4), "Artificial Intelligence" (4), and the "ACM Symposium on Principles of Programming Languages" (3).

YetRank vs. Citation Counts

The outliers in the Figure 6.3(d) have very different properties. Papers that have relatively high citation counts but a fairly low rank are referenced predominantly (97% of the time) by non-domain papers. Since YetRank includes the Impact Factor of the venues in its computations, papers that have a lot of references from non-domain papers will obtain a lower rank since the referencing papers will be from venues that have very low Impact Factor values.

The average publication year of the papers associated with the top outliers is 1993. These outliers are papers that have a relatively low rank but high citation counts. Compared to the average publication year of 1975 of the top outliers in NewRank, as expected, the rankings of papers according to YetRank depend more on the associated Impact Factor of the venues than the papers' publication dates.

Summary of Scatter Plot Analysis

Table 6.3 summarises the properties of the bottom and top outliers of the scatter plots in Figure 6.3. For each algorithm the outliers' average citation counts, publication years, impact factors and percentage of references from non-domain papers are given.

Table 6.3: Summary of the properties of the outliers in the scatter plots in Figure 6.3. The table is organised into two parts. The column "Bottom Outliers" shows the properties of the red outliers which are papers that obtained relatively high ranks according to the associated ranking algorithm but low citation counts. Conversely, the column "Top Outliers" displays the properties of the green outliers. These outliers are papers that have a high citation count compared to a relatively low rank. The columns "Cites" and "Year" show the average citation counts and publication years of the outliers. "IF" shows the average Impact Factor of the venues at which the outlier papers are published. Lastly, the column "ND" lists the percentage of references to the outlier papers that originate from papers that do not belong to the CS domain.

	Bottom Outliers				Top Outliers			
Algorithm	**Cites**	**Year**	**IF**	**ND**	**Cites**	**Year**	**IF**	**ND**
PageRank	280	1985	0.69	70.49	705	2003	3.69	4.71
SceasRank	48	1989	0.77	64.48	624	1999	2.88	2.67
NewRank	170	1998	1.37	57.35	1280	1975	0.84	9.59
YetRank	136	1983	0.69	11.42	1224	1993	1.05	97.38

It should be noted that YetRank has an advantage in ranking papers that lie at the edge of a domain, since these papers might achieve a high rank using other algorithms but by incorporating the Impact Factor of venues, YetRank has an additional factor to

rank these papers lower. This property is very helpful if only the papers that should be ranked high are in fact integral to the domain over which the algorithm is computed.

6.1.4 Score Distribution over Publication Dates

In order to further understand how the algorithms rank papers in a citation network, one can look at the score distribution over the publication years of the papers. Figure 6.4 shows the average score that papers received plotted agains the publication years.

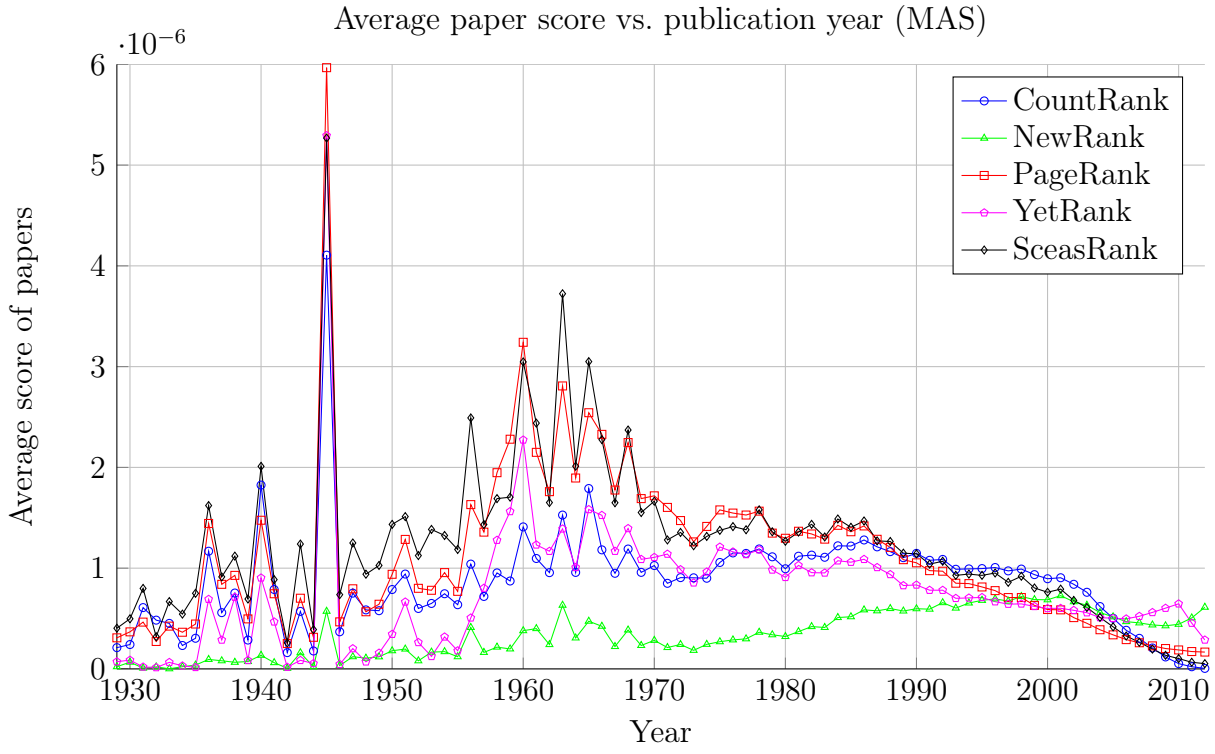


Figure 6.4: Average ranking scores of papers distributed over publication years by the various algorithms on the MAS CS citation network. The parameters of the ranking algorithms were initialised with the default values.

From the graph in Figure 6.4 one can clearly see that NewRank, compared to the other algorithms, favours newer papers. Older papers, especially those published before 1992, receive far smaller scores according to NewRank. This trend, and the sharp increase of the average score in the last two years, is due to the relatively small τ value of 4.0 that was chosen for the characteristic decay time. The highest average scores achieved with NewRank are papers published between 2000 and 2001. This means that for a τ value of 4.0, NewRank assigns highest ranks to papers that are roughly 13 years old for this particular citation network. See Section 7.1 to see how varying α and τ parameters affect the average score distribution over the publication years.

The scores assigned to new papers by CountRank tend towards zero more quickly since, on average, these papers have not been around long enough to have received a fair amount of citations.

PageRank and SceasRank focus on older papers, ranking papers published between 1950 and 1986 higher than the other algorithms. For PageRank this was to be expected as hypothesised in Section 4.2.1. This ageing effect of the PageRank algorithm can be

controlled, to some degree, by the damping factor which is shown in Section 7.1 where the PageRank algorithm is optimised for the underlying CS citation network. From Figure 6.4 one can see the similarity between SceasRank and PageRank. However, it should be noted that the default damping factor of both algorithms is 0.85 and therefore it is expected that the score distributions of the algorithms are similar.

Interestingly, the YetRank algorithm assigns larger scores to newer papers compared to the other algorithms but from 2010 to 2013 the average scores of papers decrease quickly. This is due to the Impact Factor of the associated venue which contributes to the initialisation score of a paper but is also used during the computation of the YetRank algorithm. Papers published after 2010 have barely received any citations (see the average CountRank scores in the graph) in this data set, and therefore the journals have Impact Factors of close to zero for those years, which in turn is transferred to the individual paper scores. More precisely, the Impact Factor for a venue for 2012 depends on how many citations originate from papers published in 2012 and cite papers published at that venue in the previous 5 years. Referring to Figure 5.6 one can see that the number of citations that are produced in a year drastically decreases even before 2010 resulting in relatively low impact factors for venues.

For the years 1945 and 1963 the graphs show outliers in the average scores of papers. In both cases this is due to a small number of highly cited papers that skew the results dramatically. For example, of the 20 papers published in 1945 and a total number of 1060 citations, a single paper received 921 citations which contributes 89.16% of the average score in that year according to PageRank and 76.83% by the NewRank algorithm. Similarly, in 1960 the relatively large increase in the average score for all algorithms is due to three papers who all have an above average number of citations (1543, 952, 434) compared to the average citation count of 18.19 for all 541 papers published in that year. These top three cited papers alone contribute 39.84% and 22.49% to the overall scores for that year according to NewRank and PageRank, respectively.

A similar trend can be observed in Figure 6.5 when using the DBLP data set and plotting the average ranking scores of papers against publication years. Again, PageRank assigns higher scores to older papers, while NewRank and YetRank give more focus to the more recent end of the citation network.

The outlier that can be observed in Figure 6.5 for 1960 is due to similar reasons as the outliers using the MAS data set. There are 51 papers with a total of 812 citations published in 1960. The three most cited papers have 345, 169 and 75 citations which are 72.54% of all citations received in 1960. These three papers produce 61.58% of the total scores for that year according to PageRank. Similarly, NewRank assigns 68.98% of the total scores to these papers. In the following section, details on how the algorithms handle highly cited papers are given.

It is worth mentioning that NewRank normalises the ranking scores of papers the most over time by having the smallest changes in average scores over all years. Assuming that the average quality of research output stays constant and doesn't change over the years, the average paper scores for each year should, ideally, be the same for each year. This assumption is difficult to make since all algorithms are still based on the number of citations a paper receives which in turn depends on the citation potential of a paper which reaches its maximum only a few years after its publication (see Figure 5.8 in Section 5.4). Nonetheless, assuming that the number of papers that do not receive any citations is proportional to the overall research output for each year and that the citation potential is independent of the publication date, the average scores should be roughly the same for

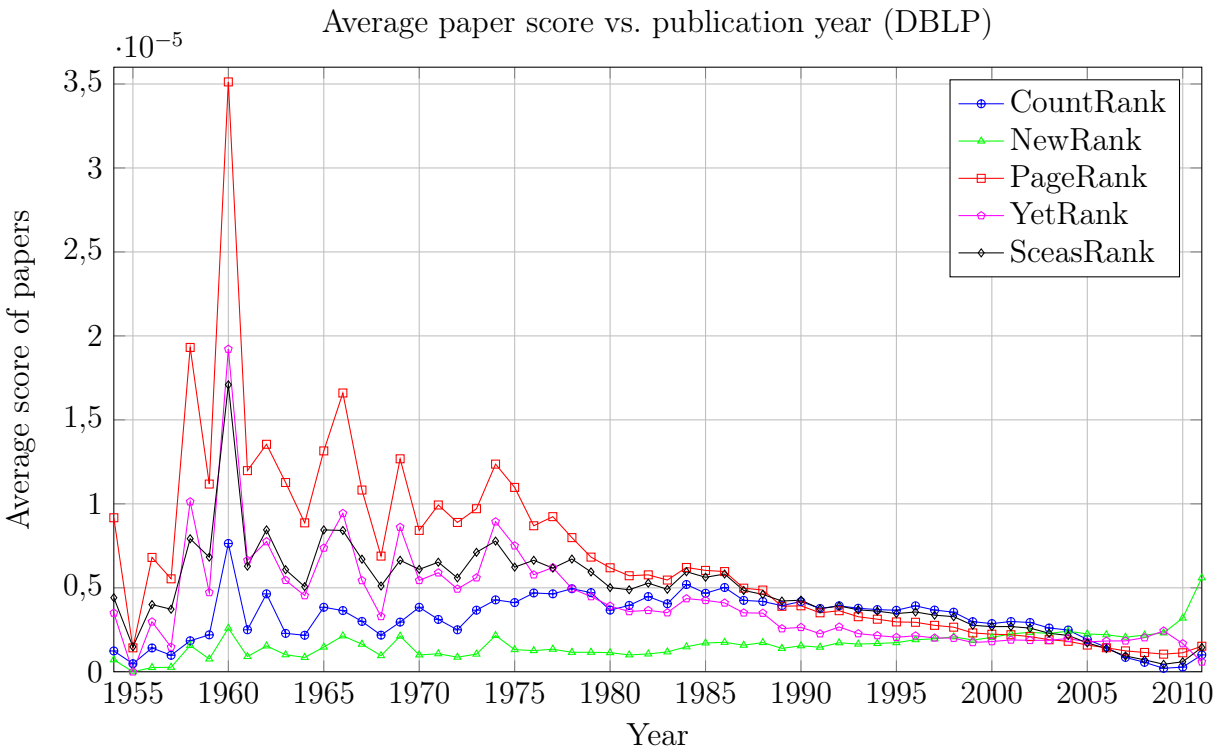


Figure 6.5: Average ranking scores of papers distributed over publication years by the various algorithms on the DBLP citation network. The parameters of the ranking algorithms were initialised with the default values.

each year.

In conclusion, when looking at the score distribution over the years, SceasRank is the algorithms whose output approximates citation counts the closest and both NewRank and YetRank focus on more recently published papers, as expected.

Top Papers Trend

Since, in general, one is interested in the top papers, the remainder of this section looks into how the top papers are ranked by the different ranking algorithms and to identify differences between the algorithms. In the previous section two outlier years in the average ranking scores were identified that were caused by papers with unusually high citation counts. On the CS citation network, for example, the outlier in 1945 contributed more towards the average score according to PageRank compared to NewRank but in 1963 the situation is switched around.

Figure 6.6 plots the contribution that the top 10% of papers for each year have on the yearly average scores. Therefore, a value closer to 100% indicates that an algorithm focuses more on highly cited papers and that less cited papers receive a smaller fraction of the yearly ranking scores.

The further one goes back in time, the closer PageRank resembles CountRank meaning that PageRank treats highly cited papers the same way as simply counting the papers' citations.

Furthermore, the two algorithms that use a characteristic decay time (NewRank and YetRank), focus much more on the top 10% of papers compared to CountRank, PageRank and SceasRank with a persistent difference of about 20%.

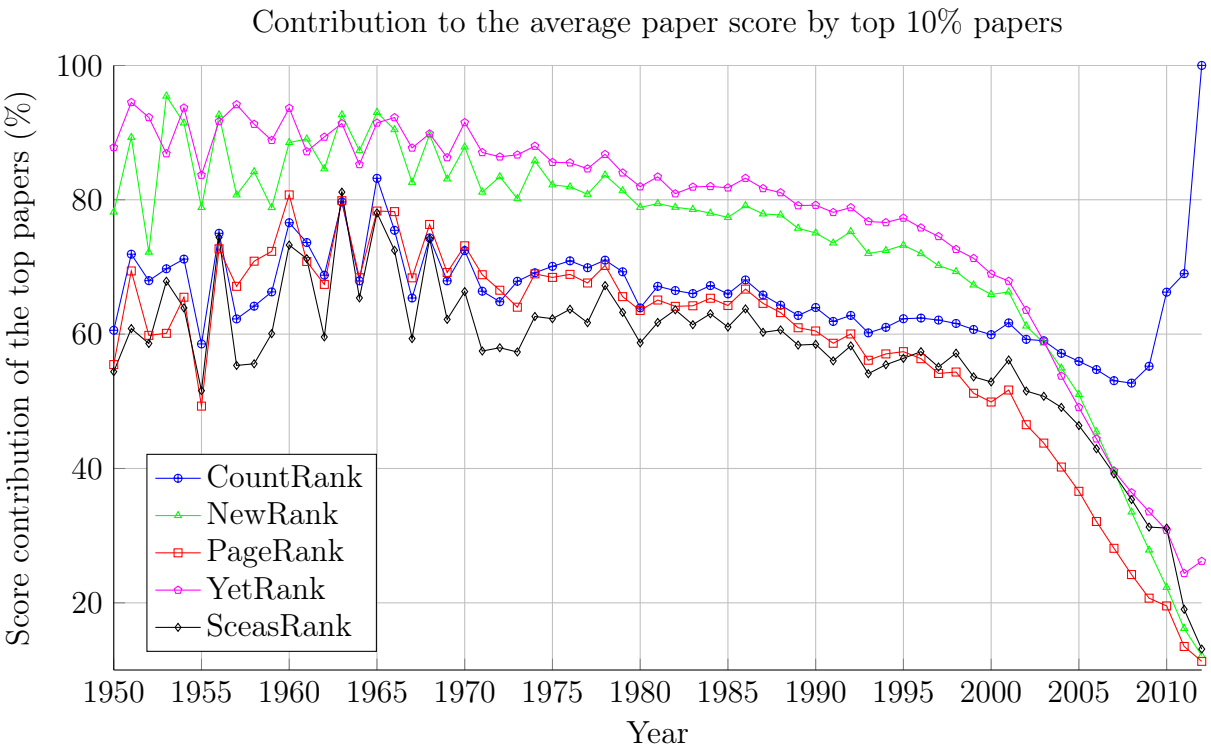


Figure 6.6: Percentage of the average score that is contributed by the top 10% of papers per publication year.

Looking at the output of the CountRank values for the most recent three years, one can see that the values are unusually high and reach 100% in 2012. This is due to the fact that only a small number of papers received any citations at all. For example, of the 2732 papers published in 2012, only 141 papers received 1 or more citations and therefore only the top 10% papers have a non-zero CountRank score.

6.1.5 Overall Top Papers

In this section the properties of the top 10 papers are discussed. For complete listings of the top 10 papers as ranked by the various ranking algorithms see Tables A.5 through A.8 in Appendix A.3. Table 6.4 shows the top 10 most cited papers in the MAS CS data set and the corresponding ranks assigned by the ranking algorithms.

The average number of citations per paper for those 10 papers is 5959.4 and the average publication year is 1991.3. Note that YetRank ranks four of these most cited papers very low. The paper "MODELTEST: testing the model of DNA substitution" is ranked at position 12 533 according to YetRank, which is an extreme outlier compared to PageRank and SceasRank which both rank this paper at position 1. This paper, in addition to the papers ranked in position 7 and 8, are published at the "Bioinformatics" journals for which the Impact Factor in 1998, 2001 and 2003 is 1.22, 2.50 and 5.39 which is relatively high compared to the average Impact Factor of 1.63 for the venues associated with the top papers listed in this table. However, these papers are ranked low by YetRank since they are cited often by papers that fall outside of the CS domain and have very low venue impact factors associated with them.

It should be noted that manuals for popular software programs are highly cited and are also highly ranked by most algorithms. The only algorithms which ranks manuals lower

Table 6.4: Top 10 most cited papers and their ranks according to the various algorithms.

	Title	Cites	Year	PR	NR	YR	SR
1	Fuzzy Sets	8954	1965	9	61	6	11
2	MODELTEST: testing the model of DNA substitution	8234	1998	1	8	12533	1
3	Matrix Computations	7822	1986	4	21	2	6
4	MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment	5875	2004	2	1	5814	2
5	Optimization by Simulated Annealing	5872	1983	8	34	4	17
6	A mathematical theory of communication	5602	2001	6	3	14	5
7	MrBayes 3: Bayesian phylogenetic inference under mixed models	4660	2003	17	7	3160	12
8	MRBAYES: Bayesian inference of phylogenetic trees	4317	2001	5	2	1320	8
9	Distinctive Image Features from Scale-Invariant Keypoints	4157	2004	273	57	34	104
10	Applied Regression Analysis	4101	1968	3	27	59	3
	Average	5959.4	1991.3	32.8	22.1	2294.6	16.9

is YetRank. No correlation between the Impact Factor of the venues at which the papers are published and the final ranks of the papers could be identified in the top papers.

From the summary in Table 6.5 one can see that NewRank ranks more recently published papers higher with an average publication year of 2000. YetRank has the oldest set of papers in the top 10 rankings with an average publication year of 1984.

Table 6.5: Properties of the top 10 papers as ranked by the ranking algorithms. The column "Avg. Cite Age" shows the average age of the citations to the top 10 ranked papers.

Algorithm	Avg. Citations	Avg. Year	Avg. Cite Age
CountRank	5959.40	1991.30	14.14
PageRank	5225.90	1989.02	14.98
NewRank	3632.90	2000.60	6.12
YetRank	4328.40	1984.10	19.15
SceasRank	5013.80	1993.70	11.92

The average citation age of the top papers also varies significantly between the algorithms. NewRank is the only ranking algorithm that considers the age of citations in its computations. This becomes evident when comparing the top papers, since the average citation age is 6.12 years compared to over 11 years for all other algorithms.

6.1.6 Identifying Current Research Activity

Purpose

It can be argued that it is important to identify current research activity since further insight into which fields are currently popular and are actively researched can help researchers and scholars choose research topics and aid funding bodies in the decision of grant allocations.

Unfortunately, papers are not classified into granular research topics which makes it difficult to identify current research trends in terms of topics. Nonetheless, the entire research trend can still be used to evaluate the algorithms in identifying current research activity. It can be assumed that the most recently published papers constitute the current research performed. Therefore, the citations of these papers can be used as a measure for current relevance for the referenced papers and their importance to current research interests.

Research Method

From the MAS CS data set the papers published between 2010 and 2013 are selected and their citations to papers published in or before 2009 used to evaluate the ranking algorithms in identifying current research activity. In other words, young papers published between 2010 and 2013 are pruned from the citation network. The ranking algorithms are computed over a citation network constructed from the remaining papers that are published in or before 2009 which now contains 2 117 390 papers and 11 183 776 references.

The set of young papers constitutes 11.6% of all CS papers and produce 1 660 506 citations referencing papers in the citation network over which the ranking algorithms are computed.

The rankings of the algorithms on the subset of papers is compared to the citation counts accrued from the set of recently published papers, using the Pearson correlation coefficient. The Pearson correlation coefficient is used since the citation counts of papers are compared to their ranking scores. These results are given in Table 6.6.

Results

From Table 6.6 one can see that simply using citation counts predicts the current research activity the most accurately with a correlation of 0.729. Using the default values of the algorithms, SceasRank performs the best (0.644) followed by NewRank (0.597) and PageRank (0.561).

The parameters of the algorithms can be fine-tuned to find optimal parameters that achieve higher correlation with the citation counts of the recent papers. From Table 6.6 one can see that NewRank, if used with $\alpha = 0.35$ and $\tau = 16.0$, achieves the highest correlation of 0.669 but is still 0.06 points below CountRank.

Table 6.6: Pearson correlation values r between the number of citations accrued by papers in recent years and the ranking results of the algorithms on the MAS CS citation network. The results using the algorithms' default parameters are given on the left. On the right, the parameter values for which the highest correlation is achieved are given for each algorithm.

Algorithm	**Default Parameters**	r	**Optimal Parameters**	r
CountRank	None	0.729	–	–
PageRank	$\alpha = 0.85$	0.561	$\alpha = 0.25$	0.644
NewRank	$\alpha = 0.85$, $\tau = 4.0$	0.597	$\alpha = 0.35$, $\tau = 16.0$	0.669
YetRank	$\alpha = 0.85$, $\tau = 4.0$	0.503	$\alpha = 0.55$, $\tau = 2.0$	0.636
SceasRank	$\alpha = 0.85$, $a = e$, $b = 1.0$	0.644	$\alpha = 1.0$, $a = 3.5$, $b = 0$	0.644
SceasRank1	$\alpha = 1.0$, $a = e$, $b = 1.0$	0.643	–	–
SceasRank2	$\alpha = 0.85$, $a = e$, $b = 0$	0.644	–	–

It is interesting to note, that the parameter b of SceasRank does not have an impact on the correlation between the rankings of the papers and their citation counts from recent papers, given that $\alpha \in [0, 1)$. If $\alpha = 1$, then b has to be greater than 0 to obtain even moderate correlation. Nonetheless, if $\alpha = 1$ and $b > 0$, the correlation is only dependent on the values of α and a. It should be noted that high correlations are found for SceasRank when $\frac{\alpha}{a} = [0.22, 0.32]$ independent of the values of b. Furthermore, the results do not depend on using a modified citation network where edges are added to the dangling vertices (see method (2) Section 2.4.1) or using an unmodified citation network.

6.2 Comparing Venue Ranking Algorithms

The venue ranking algorithms compared in this section are the Eigenfactor method, the Impact Factor and the h-index. Since both the Eigenfactor and the Impact Factor methods use the idea of a census and a target window, the methods are compared using the same census and target windows of $[2009; 2009]$ and $[2004; 2008]$, respectively. The year 2009 is chosen for the census window since the MAS CS citation network contains the most published papers in that year. Moreover, 2009 is the year in which most references are produced. The Impact Factor's target window size is 2 years by default. In order to compare the results to the Eigenfactor metric, a larger target window size of 5 years is chosen which is the default target window size of the Eigenfactor metric. The same is done for the computations of the citation counts and the h-indices of venues. Only citations that originate from papers in the census window and reference papers in the target window are considered. The CountRank method for venues simply counts the total number of citations that papers published at the venues during the target window receive. The citation counts are not normalised by the number of papers published during the target window since this would essentially be the Impact Factor metric. The damping factor for the Eigenfactor metric is set to the default value of 0.85.

6.2.1 Correlations between Venue Ranking Algorithms

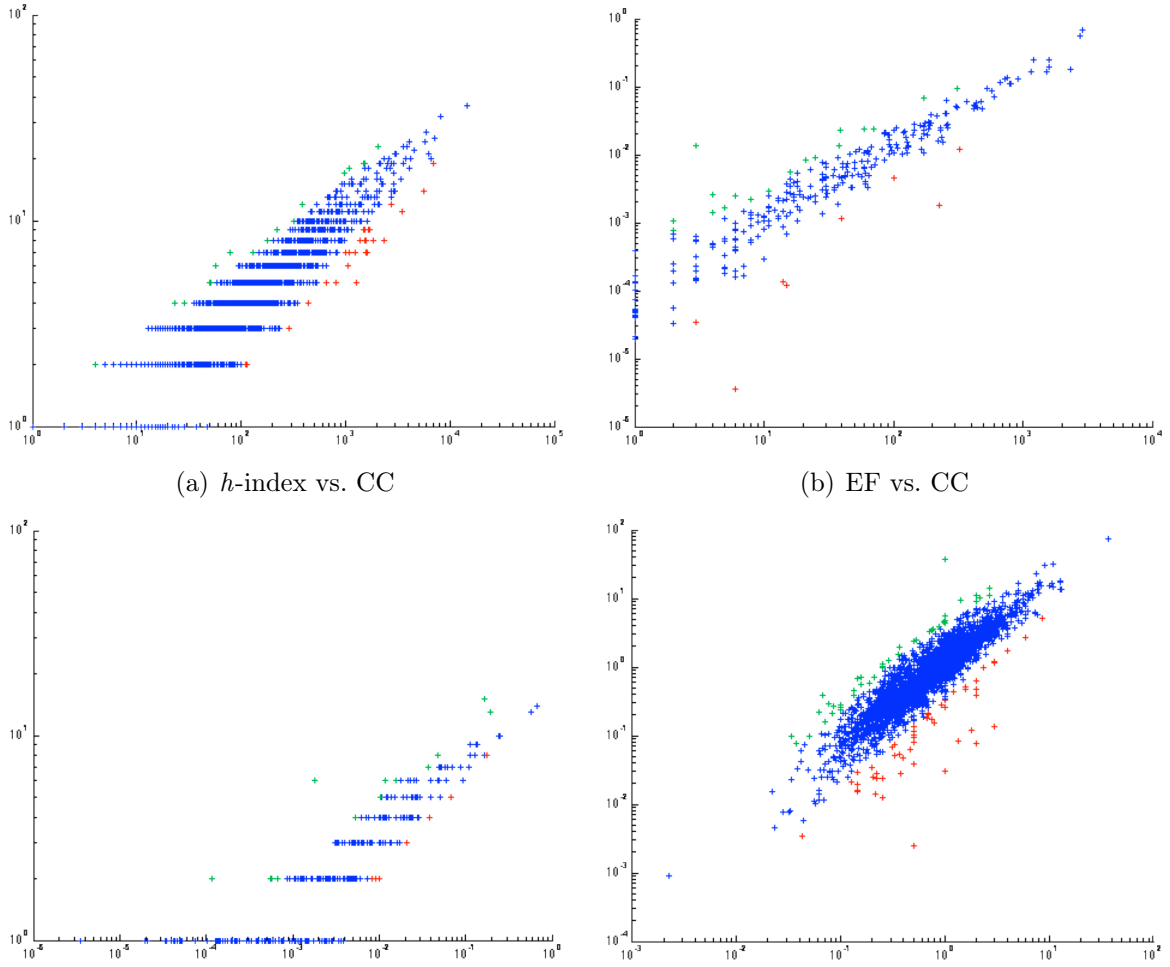
CountRank (CCR) and the Eigenfactor (EF) metric compute overall importance scores of venues. Alternatively, the Article Influence (AI) score of the Eigenfactor metric and the Impact Factor (IF) calculate a per-article prestige score for venues and therefore are not compared to the previously mentioned methods. The h-index cannot be classified into one of the two groups since both the notion of a venue's overall impact and the individual influence of its papers is incorporated into the score. The results of the h-index are therefore compared to all metrics in the following sections.

Table 6.7: Spearman correlation coefficients (ρ) between the rankings of all venues of the CS domain for each pair of algorithms where applicable. The highlighted cell indicates the highest correlation while the boxed cell shows the lowest correlation between two algorithms.

	IF	***h*-index**	**EF**	**AI**
CCR	N/A	0.742	0.962	N/A
IF	–	0.763	N/A	0.898
***h*-index**	–	–	0.736	0.642
EF	–	–	–	N/A

Table 6.7 shows the Spearman correlation values of comparing the output values of the different venue ranking metrics. The h-index has a higher correlation with the EF metric than with the AI scores. If the Eigenfactor metrics is considered the gold standard, then it appears that the h-index computes a score that is closer to a venue's overall importance to the scientific community than the average influence that the venue's papers have.

6.2.2 Comparison using Scatter Plots



(a) h-index vs. CC

(b) EF vs. CC

(c) h-index vs. EF

(d) AI vs. IF

Figure 6.7: Scatter plots of the ranks of venues for different venue ranking metrics. The h-index, the Eigenfactor Metric (EF), the Journal Impact Factor (IF), and the Article Influence (AI) of the Eigenfactor Metric are considered. For each plot, the red and green data points indicate outliers with high and low x/y ratios, respectively.

Figure 6.7(a) plots the h-index of venues against the total citation counts that papers, published at the corresponding venues, have received. Similar to the scatter plot analysis in Section 6.1.3, venues that are outliers are highlighted in different colours which are used for further analyses. The red outliers to the right of the main scatter plot body are papers that received relatively high citation counts compared to a relatively low h-index value. Alternatively, the green data points indicate papers that are outliers since they receive a high h-index value with a comparatively low total citation count.

When comparing the venues associated with the red and green outliers in Figure 6.7(a) not many differences are identified. The only differences between the outliers are that the venues associated with the red outliers have a much larger paper count (2459 on average) but a low average citation count of 17.69. Alternatively, the green outliers are venues that contain a small number of papers (78 on average) but have a high citation count of 86.67 on average. The venues associated with the green outliers seem to be more selective.

Considering the scatter plot in Figure 6.7(b) the largest differences between the outliers is that the green outliers are venues with a low self-citation rate of 6.54%. Comparatively, the venues associated with the red outliers have a high self-citation rate of 81.04%. A similar trend is exhibited by the scatter plot in Figure 6.7(d), where the venues associated with the green outliers have a low self-citation rate of 2.88% compared to 65.09% for the red outliers.

6.3 Comparing Author Ranking Algorithms

In this section the algorithms that can be used to rank authors are compared. The input for the ranking algorithms is the MAS CS citation network. CountRank (CCR) simply counts the number of citations that authors have received for their published papers excluding author self-citations. The results of using CountRank with author self-citations included (CCRS) are also given for comparison. The α parameter for the Author-Level Eigenfactor (AF) method is set to the default value of 0.85 and author self-citations are omitted. For the h-index, g-index and *i10*-index methods author self-citations are included. In addition, rankings according to publication counts (PC) of authors are also given. In Section 6.3.1 the similarity of the ranking algorithms is analysed using correlation coefficients. Scatter plots are used in Section 6.3.2 to compare the ranking outputs of the various algorithms.

6.3.1 Correlation between Author Ranking Algorithms

The top 50 ranked authors according to each algorithm are used and the number of common authors counted for each pair of metrics. The results are listed in Table 6.8.

Table 6.8: Number of common authors in the top 50 rankings of each pair of author ranking algorithms.

	CCRS	**AF**	***h*-index**	***g*-index**	***i10*-index**	**PC**
CCR	46	13	32	41	24	9
CCRS	–	14	32	43	26	9
AF	–	–	10	15	7	2
***h*-index**	–	–	–	31	33	12
***g*-index**	–	–	–	–	23	7
***i10*-index**	–	–	–	–	–	18

The largest number of common authors are in the rankings produced by CCR and CCRS. This is expected since both metrics count the number of citations that authors receive except that CCRS includes author self-citations. The most similar metric to pure citation counts (CCRS) is the g-index with 43 authors in common in the top 50 ranks. Note that all metrics have more common authors in their rankings when compared to

CountRank with self-citations than compared to CountRank in which self-citations are omitted. This is expected since all metrics include self-citations by default. The only metric that excludes self-citations by default, namely AF, also has a higher number of common authors with CCRS.

Comparing the algorithms by only looking at the top 50 ranked authors does not give a full picture about the similarity of the various metrics. Therefore, Table 6.9 lists the Kendall rank correlation coefficient values on the complete rankings for each pair of metrics.

Table 6.9: Kendall rank correlation coefficients (τ) for the complete author rankings of the CS domain for each pair of algorithms. Highlighted cells indicate a high correlation while the boxed cell shows the lowest correlation between two algorithms.

	CCRS	**AF**	***h*-index**	***g*-index**	***i10*-index**	**PC**
CCR	0.960	0.734	0.617	0.734	0.517	0.450
CCRS	–	0.747	0.634	0.754	0.507	0.483
AF	–	–	0.594	0.651	0.451	0.438
***h*-index**	–	–	–	0.659	0.371	0.519
***g*-index**	–	–	–	–	0.542	0.670
***i10*-index**	–	–	–	–	–	0.474

When ignoring the high correlation between CCRS and CCR, the highest correlation according to the Kendall's τ correlation coefficient is found between CCRS and the g-index with a value of 0.754. Considering the publication counts (PC) of authors, the g-index is the most similar while AF has the lowest correlation. Comparing CCR and CCRS with the other metrics, citation counts including self-citations (CCRS) always has a higher correlation than CCR. This is to be expected since all metrics except AF include author self-citations by default. As seen before with the number of common authors in the top 50 rankings, AF also has a higher correlation with CCRS, even though AF excludes self-citations.

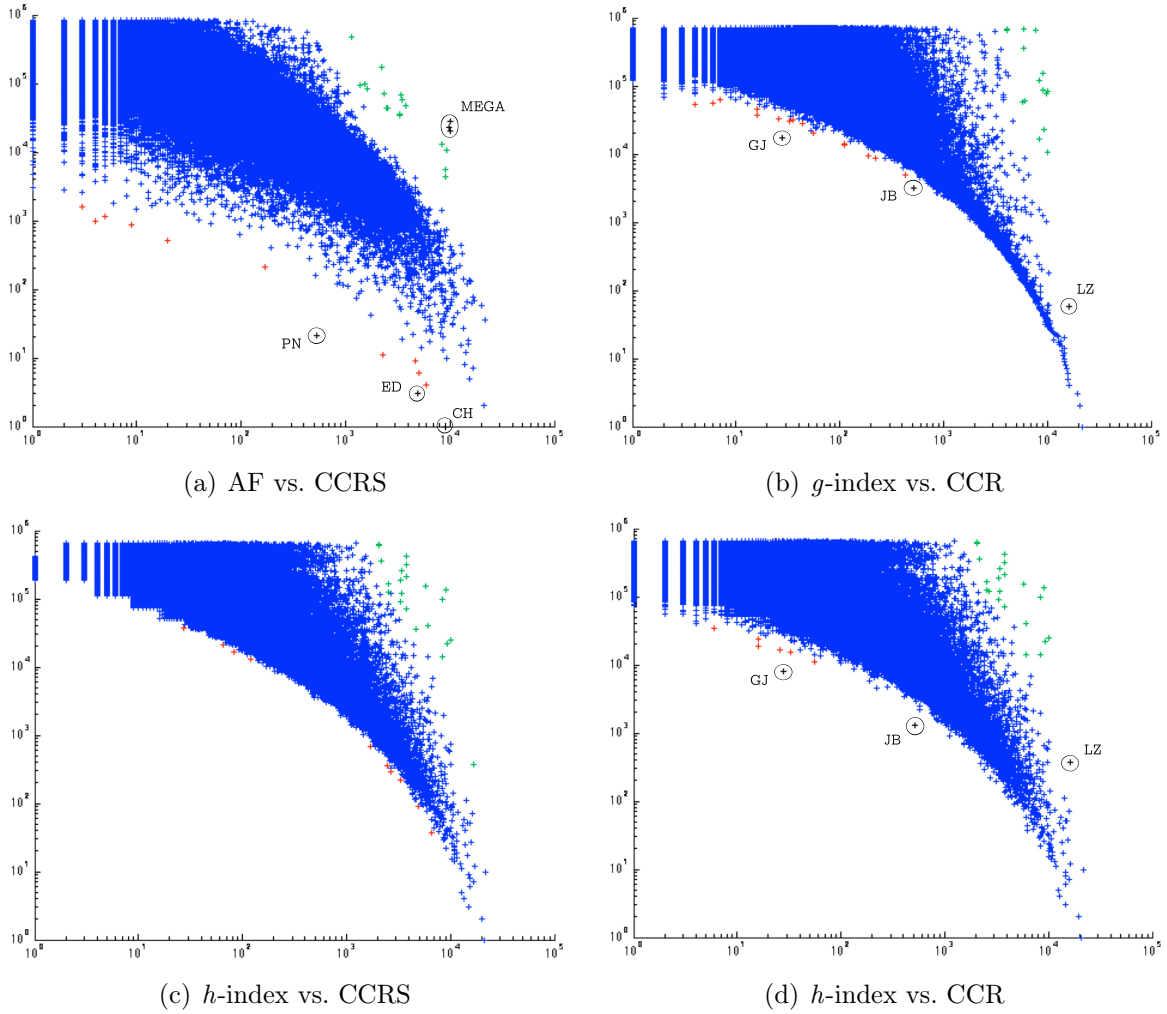
The top 10 ranked authors by the AF method are listed in Table A.9 in Appendix A.3 with the corresponding ranks as computed by CCRS, the h-index, the g-index and the $i10$-index.

6.3.2 Comparison using Scatter Plots

A similar approach is used to compare the author ranking algorithms using scatter plots as previously done for the paper and venue ranking algorithms. Again, certain outliers are highlighted in different colours to gain more information about how the ranking algorithms rank authors.

Author-Level Eigenfactor vs. Citation Counts

The first figure (6.8(a)) plots the ranks of authors as computed by the Author-Level Eigenfactor algorithm against their citation counts including self-citations. The bottom outliers (red) are 13 authors that are ranked relatively high, on average 413, but have comparatively few citation counts with 54 citations per paper on average. The top outliers (green) are 24 authors that have a relatively high number of citations but are ranked low



(a) AF vs. CCRS

(b) g-index vs. CCR

(c) h-index vs. CCRS

(d) h-index vs. CCR

Figure 6.8: Authors' ranks according to the Author-Level Eigenfactor (AF) metric, their h-index and their g-index plotted against their citation counts, with self-citations included (CCRS) or omitted (CCR). The y-axis indicates the ranks according to the various metrics with the first rank at the bottom. The x-axis indicates the citation counts of the authors with the highest citation counts on the right. The red and green data points indicate outliers with low and high rank to citation count ratios, respectively. The black data points are far outliers that are used to obtain further insight into the ranking properties of the algorithms.

by AF. The average rank of the green outliers is 84 773 while the average citation count per paper is 641.

The average publication year of all the papers published by the authors that are associated with the top outliers is 2001. Of all the citations citing these papers 98.79% are references from papers that do not belong to the CS domain and 0.08% are author self-citations. The average number of collaborators that the authors have worked with is 15 for the top outliers.

Considering the bottom outliers, the average publication year is 1981, 5.17% are citations from non-domain papers, 1.03% are author self-citations, and the average number of collaborators per author is 22.

Besides the average number of citations per paper that the authors receive, the biggest difference between the top and bottom outliers is that the authors that fall into the top outliers receive many citations from non-domain papers.

The three black outliers at the bottom are authors that have published a lot of relatively highly cited papers a long time ago. The author Charles Hoare (CH) published 153 papers with an average publication year of 1986 and an average of 59 citations. Similarly, Edsger Dijkstra (ED) published 68 papers with an average publication date of 1979 which have obtained 73 citations on average. Lastly, Peter Naur (PN) published 56 papers which have an average publication year of 1970 and received 10 citations on average. Note that all of these authors received the prestigious Turing Award [69].

Considering the other three black outliers at the top of the figure, the authors have published relatively few papers (10 on average) in the CS domain but have received a lot citations (9970 on average). Most of these citations (99.91%) are citations from non-domain papers. The three authors (MEGA in Figure 6.8(a)) are the developers of the "MEGA: Molecular Evolutionary Genetics Analysis" software program used for statistical analysis of molecular evolution. Masatoshi Nei led the development of this software with his graduate student Sudhir Kumar and postdoctoral fellow Koichiro Tamura [70]. All three published the papers that are cited when using different iterations of the software product and therefore gaining a lot of citations. The papers for version 2 and 3 are both in the top 20 papers according to citation counts and ranked at positions 19 and 4, respectively.

Based on these outliers, the AF methods seems to distinguish between authors with an overall significant impact to science from authors that are more of "one-hit wonders", authors that published a few papers that gained a lot of citations. Authors with low ranks according to the AF metric have a low scientific output (7 papers on average), collaborate with only a few (15 on average), but have very high citation rates with 641 citations on average.

On the other hand, authors that are ranked high by the AF metric are authors that contribute a lot to the scientific enterprise with an average of 47 papers published and are intrinsic to the field of Computer Science since the average percentage of references from non-domain papers is only 5.17%.

It should be noted that the AF metric computes overall scores for the importance or the level of contribution of authors and may not pick up authors that have only recently started their careers. Of the bottom outliers, for example, the average publication year of papers, published by the associated authors, is 1981 which is relatively old.

The g-index vs. Citation Counts

Figure 6.8(b) shows the ranks of authors according to the g-index plotted against their citation counts with self-citations excluded. The bottom outliers (green) are 17 authors with a large number of publications (68 on average) and a very high self-citation rate of 70.19%. Their papers are cited rarely with 5 citations on average. The top outliers are 17 authors that, on average, published 11 papers that are cited very often with an average citation count of 732 and a self-citation rate of 0.13%. Another property that differs between the bottom and top outliers is that the authors associated with the bottom outliers published papers that are rarely cited by non-domain papers (4.68%) compared to the top outliers (72.14%).

The two black outliers at the bottom of the plot are Giorgi Japaridze (GJ) and Jan Bergstra (JB). Both of these authors a have a very high self-citation ratio. GJ and JB have published 120 and 30 papers respectively with a self-citation rate of 62.87% and 91.08% but have a low citation count of 12 and 10 on average.

The third highlighted author is Lotfi Zadeh (LZ) who wrote the most cited paper "Fuzzy Sets" with 8 954 citations alone. In total, LZ obtained 16 481 citations for his 92 papers in this data set with a self-citation rate of 0.74%.

From the distribution of this scatter plot one can see that the similarity between an author's g-index and the number of citations becomes more apparent the more papers an author has published and the more citations they have received.

The h-index vs. Citation Counts

The ranks according to the h-index metric are plotted twice. Figure 6.8(c) shows the h-index plotted against the citation counts of authors with self-citations included. In Figure 6.8(d) the ranks of authors according to the h-index are plotted against the citations counts of authors where self-citations are omitted.

The h-index includes self-citations by default and the two scatter plots are used to depict the difference when the h-index is compared to CCRS and CCS. From Figure 6.8(d) one can see that the bottom outliers (red) are further away from the main scatter plot body when the h-index is compared to citation counts with self-citations are omitted. This is to be expected since the h-index includes self-citations and should have a higher correlation with citations counts that include self-citations. This is also true for the g-index. However, it should be noted that the scatter plots of AF plotted against CCRS and CCR have no noticeable differences and therefore AF plotted against CCR is omitted.

The outliers in Figure 6.8(d) are very similar to the outliers in the corresponding plot for the g-index. Again, the bottom outliers are authors that have a high percentage of self-citations with 75.46% of all references to their papers being self-citations. The top outliers are authors that published papers that received a lot of references from non-CS papers with 67.20% on average. The far outliers highlighted in black are the same authors that are also far outliers in Figure 6.8(d).

6.4 Chapter Summary

It is very difficult to compare different ranking algorithms to each other on real academic citation data. Firstly, there does not exist any ground truth about the quality, significance or impact of various academic entities. Secondly, the size of any real-world data set is too large for manual evaluation.

This chapter covered a number of different ways of comparing the ranking algorithms to identify certain characteristics. Most comparisons discussed in Sections 6.1 through 6.3 are analyses that compare only specific properties of the algorithms and cannot be seen as holistic evaluations. The following main properties were identified:

- Of all PageRank-like algorithms, SceasRank converges the fastest.
- The time and space complexity of CiteRank renders the computation on the MAS CS citation network infeasible.
- SceasRank and PageRank exhibit the most similar behaviours.
- Compared to the other algorithms, YetRank does not rank papers with many references from non-domain papers very highly.

- With the exception of YetRank, the algorithms assign high ranks to manuals of popular software programs.
- The highest dependance on the publication dates of papers is NewRank by focusing the most on recently published papers.
- When using the MAS CS citation network the edge of the network (specifically non-CS papers referencing CS papers) has a large influence on the results of the ranking algorithms. This is due to the fact that 35.24% of all citations in the network originate from non-CS papers. YetRank appears to be affected the least by this behaviour since venues of non-CS papers have close to zero Impact Factors which are incorporated in YetRank's computation.
- By comparing the author ranking metrics, the g-index has a closer overall correlation to an author's number of citations than the h-index.

When using the citations from recently published papers as a measure of current research activity, then simply counting citations is the best metric evaluated to identify current important papers. Since the evaluation data for this experiment is based on pure citation counts, this result was expected. The algorithm that identifies current research activity the best apart from CountRank is NewRank.

It should be noted that most of the above-mentioned properties might depend on the underlying citation network and cannot be assumed valid in general. Furthermore, all the experiments in this section were conducted by using the default parameter values of the algorithms. The properties can change if the parameters are varied.

Chapter 7

Evaluating Ranking Algorithms

This chapter presents empirical results that are obtained by evaluating the performance of the algorithms against four different test data sets that are based on expert opinions.

Firstly, a list of 207 academic papers that have received accolades as important and high-impact publications was collected. This list was obtained for 14 different conferences that usually hand out the prizes 10 to 12 years after the papers' initial publications.

Secondly, a set of 372 papers that won best paper awards at CS conferences was identified to evaluate the precision of award committees from 29 conferences in identifying important papers at the time of publication.

In addition, a data set of authors that have won prizes in recognition of their innovative and long-lasting contribution to science was compiled to evaluate the author ranking algorithms.

Lastly, a list of important papers that are regarded as breakthrough publications by significantly changing scientific knowledge was collected for Computer Science. This set of papers is used to evaluate how well the paper ranking algorithms identify overall important papers.

7.1 Evaluating Paper Ranking Algorithms

Purpose

It is important to evaluate the ranking algorithms on their precision and recall performances on identifying high-impact papers using an independent criterion. This is done by using the set of *award papers* which are papers that won prizes for being most influential in their fields and yielded high impact. The prizes are selected by reviewing panels of the various venues and therefore can be assumed to be picked by experts in their fields. Using the same set of papers the algorithms are optimised in order to find the optimal parameters for the algorithms.

Experimental Method

The set of award papers consists of 14 lists from different conferences and in total encompasses 207 papers which were awarded a high-impact paper prize.

Since the award papers all belong to the CS domain, only the subset of CS papers from the MAS data set are used as input for the ranking algorithms. Therefore, the citation network used consists of 2 394 976 papers, 12 907 440 citations, 1 351 journals, and 3 152 conferences.

Even though the computations are only executed on the CS citation network, it is is relatively large and therefore infeasible to evaluate the performance of the CiteRank algorithm.

For comparison reasons the DBLP data set is also used to evaluate the algorithms. The DBLP citation network with 469 940 papers and 2 083 983 references is much smaller and the evaluation of CiteRank using this data set was possible.

The PageRank algorithm is executed with a damping factor of $\alpha = 0.85$. Similarly, the NewRank algorithm is initialised with $\alpha = 0.85$ and $\tau = 4.0$. The YetRank algorithm uses a target window size of 5 years and is initialised with the parameters $\alpha = 0.85$ and $\tau = 4.0$. The default parameters for the different SceasRank variants are used as described in Section 4.2.2. For all algorithms a precision threshold of $\delta = 1.0 \times 10^{-6}$ is used.

The output of a ranking algorithm, a ranked list of papers, is evaluated using the set of award papers. As a precision metric the number of award papers that are ranked the highest within their venue and their respective publication years is used. The percentage of the number award papers ranked at position one is calculated for each venue. The average percentage of all venues is then calculated and referred to as **% Hits** in the following section.

The recall metric used is similar to the precision metric described before but instead of counting the number of award papers that are ranked at position one, the number of award papers that are ranked in the top 10 ranks is calculated. Again, the percentage of award papers that are ranked in the top 10 ranks is computed for each venue and the average percentage is used as an overall recall indicator. This recall value is referred to as **% Top 10** in the following sections. The number of top papers in which to search for is chosen to be 10 because result pages usually list 10 items per page.

Another metric is used to compare the performance of the ranking algorithms and is referred to as **Avg. Dist**. This metric calculates the average difference between the rank of an award paper to the top ranked paper within the same venue and the corresponding year. In other words, it calculates the average number of non-award papers that are ranked higher than the award papers themselves.

The **F1 Score** is the harmonic mean of the precision (**% Hits**) and recall (**% Top 10**) values. Formally, it is defined as follows:

$$\text{F1 Score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

The **average precision** is a single value that encompasses both the precision and recall accuracy of m ranked elements in a query that returns a results set of size n. It is often used in the field of information retrieval and is defined as follows:

$$\text{AP@}n = \frac{1}{\min(m, n)} \cdot \sum_{k=1}^{n} \frac{\text{P}(k) \cdot \text{rel}(k)}{k}$$

where $\text{P}(k)$ is the precision at cut-off k in the result set (described below) and $\text{rel}(k)$ is a function that returns 1 if the element with rank k is relevant and 0 otherwise. $\text{P}(k)$ is the number of relevant elements found in the first k ranked elements. For example, consider three ICSE award papers that were published in 1990 and ranked in positions 1, 3 and 11 in a list of all publication published at ICSE in 1990. The average precision (AP@10) for ICSE for 1990 would be $(\frac{1}{1} + \frac{2}{3})/3 = 0.56$.

The **mean average precision** is the mean of a set of N queries, therefore

$$\text{MAP@}n = \frac{1}{N} \cdot \sum_{i=1}^{N} \text{AP@}n(i)$$

The MAP@10 is used in the following sections and shows the average precision of the 14 sets of award papers. More precisely, the mean average precision (MAP@10) is computed for each venue where the average precision (AP@10) of each publication year of the award papers for that venue is averaged. In the following sections **AMAP@10** refers to the average MAP@10 scores over all venues.

Results

The results of evaluating the ranking algorithms using the award papers and applying the metrics described above are listed in Table 7.1.

Table 7.1: Results of evaluating the ranking algorithms using the MAS CS citation network as input against 207 high-impact award papers from 14 CS conferences. Column "% Hits" displays the percentage of award papers that scored the highest ranking values in their respective years and publication venues. The column "% Top 10" shows the percentage of award papers that were listed as one of the top 10 ranking results in their respective years. "Avg. Dist" shows the average number of papers that achieved better scores than the award papers for a certain year within the corresponding publication venue. The column, "F1 Score", displays the harmonic mean of the precision (% Hits) and recall (% In Top 10) values. A MAP@10 value is calculated for each venue and their average is shown in column "AMAP@10".

Algorithm	% Hits	% Top 10	Avg. Dist.	F1 Score	AMAP@10
CountRank	44.44	90.34	3.96	0.60	0.61
PageRank	40.58	90.34	3.97	0.56	0.57
NewRank	34.78	88.41	4.66	0.50	0.54
YetRank	42.03	91.30	3.77	0.58	0.59
SceasRank	43.96	91.30	3.72	0.59	0.60
SceasRank1	44.93	91.30	3.71	0.60	0.60
SceasRank2	43.96	91.30	3.72	0.59	0.60

From the results in Table 7.1 one can see that SceasRank1 achieves the highest precision by assigning highest ranking scores to 44.93% of the 207 award papers. The highest recall values are produced by YetRank and the SceasRank algorithms by placing 91.30% of the award papers in the top 10 ranks in their respective years and at their corresponding conferences. In addition, the average distance of the award papers to rank one of all corresponding conference papers in the respective years is also the smallest according to SceasRank1. The highest AMAP@10 is achieved by CountRank with a value of 0.61.

Since the previously mentioned results are computed on the entire CS data set of papers with publication dates ranging until 2013, it is possible that the results are skewed by citations from papers that were published after the most-influential prize was awarded. It seems reasonable to assume that after articles win a MIP award their visibility increases making them more likely to be cited in the years following the prizewinning. Table 7.2 shows the results of using the ranking algorithms on adjusted data sets that only contain publications up to the years of award consideration. Therefore, the adjusted data sets contain all the papers that were published by the time the MIP prizes were awarded. For

example, given that an award paper wins the MIP award in 2008 and was published in 1998, the input data set only contains references from papers published in or before 2008 and therefore excludes all references produced after 2008.

Table 7.2: Results of evaluating the ranking algorithms against the 207 award papers from 14 conferences using a reduced citation network of MAS CS papers.

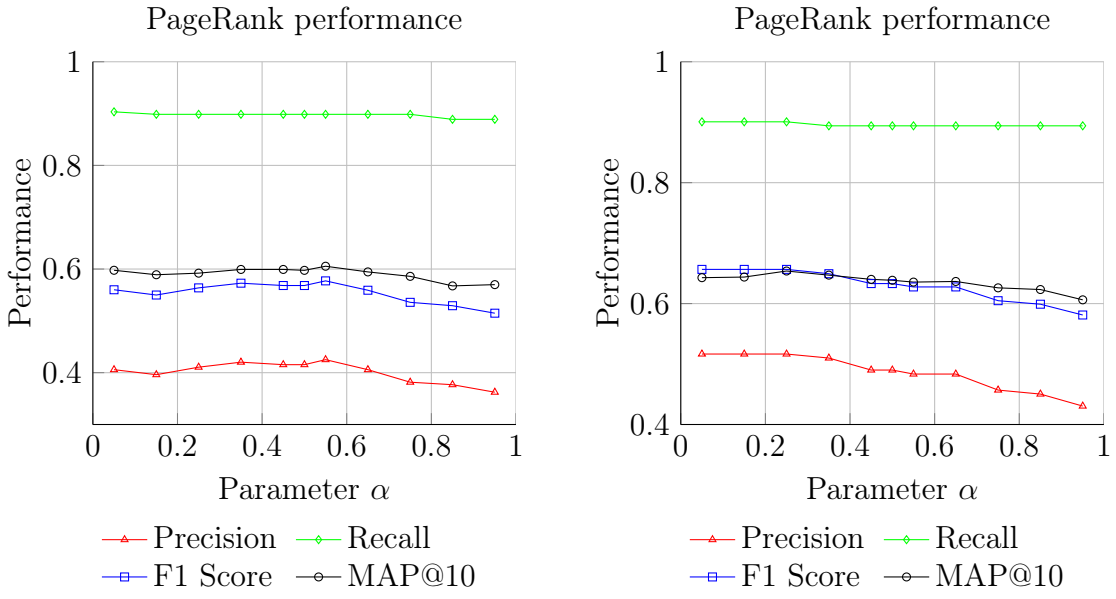
Algorithm	% Hits	% Top 10	Avg. Dist.	F1 Score	AMAP@10
CountRank	40.10	89.86	4.43	0.55	0.59
PageRank	39.13	88.89	4.12	0.54	0.56
NewRank	35.27	87.92	4.53	0.50	0.55
YetRank	43.48	89.86	3.71	0.59	0.61
SceasRank	41.06	90.34	3.98	0.56	0.58
SceasRank1	41.55	89.86	3.93	0.57	0.59
SceasRank2	41.06	90.34	3.98	0.56	0.58

By computing scores on the adjusted data sets the ranking algorithms perform differently as can be seen in Table 7.2. In terms of precision, YetRank performs the best by placing 43.10% of the award papers highest in their respective years. When comparing the change in precision only YetRank and the NewRank algorithms improve on the results of using the entire MAS CS citation network. The same is true for the AMAP@10 values. These two algorithms consider the publication dates of the papers in their computations. Both YetRank and NewRank rank recently published papers higher compared to the other algorithms and shift the average score of papers towards the more recently published papers. This might be the reason that these algorithms improve on the precision of predicting the award papers when the network is truncated. Considering the recall values, SceasRank achieves the best results, by ranking 89.86% of the award papers in the top 10 ranks in their years.

The same test data of award papers is used with the entire DBLP data set. The sample size is smaller compared to the MAS CS data set since some award papers could not be matched with the corresponding papers in the DBLP data set or do not have any citations in the data set. The award papers from the Special Interest Group on Genetic and Evolutionary Computation (Sigevo) could not be matched to the DBLP data set and all award papers from the AAAI conference have in-degrees of zero. Therefore, the resulting test data set contains 151 award papers from 12 different conferences. See Table A.2 in Appendix A.2 for a detailed listing of the award papers used.

Table 7.3: Results of evaluating the ranking algorithms using the DBLP citation network as input against 151 award papers from 12 Computer Science conferences.

Algorithm	% Hits	% Top 10	Avg. Dist.	F1 Score	AMAP@10
CountRank	56.29	92.05	2.44	0.70	0.66
PageRank	45.70	90.07	3.05	0.61	0.62
NewRank	39.07	86.09	3.75	0.54	0.51
YetRank	45.70	90.07	2.81	0.61	0.59
SceasRank	52.32	90.07	2.75	0.66	0.65
SceasRank1	51.66	89.40	2.74	0.65	0.64
SceasRank2	52.32	90.07	2.75	0.66	0.65
CiteRank	46.36	94.04	2.58	0.62	0.59



(a) Results on the MAS CS subset network.

(b) Results on the DBLP data set.

Figure 7.1: Performance of PageRank with varying α parameters on the MAS CS citation network in 7.1(a) and the DBLP data set in 7.1(b). For the MAS CS citation network, PageRank performs the best with a damping factor value of $\alpha = 0.55$, while for the DBLP data set the optimal value is $\alpha = 0.25$.

Similar results as seen before are observed when the DBLP data set is used to evaluate the algorithms as shown in Table 7.3. CountRank performs the best with a precision of 56.29% and recall value of 92.05%. Comparing the AMAP@10 values all algorithms perform in the same order except that PageRank outperforms YetRank when using the DBLP data set. In addition, the results of evaluating CiteRank over the DBLP citation network are given. CiteRank achieves the overall highest recall value of 94.04% but a fairly low precision value of 46.36% compared to the other algorithms.

Optimising PageRank

As previously described in Section 2.5.1 the damping factor α of the PageRank algorithm not only has an impact on the connectedness of the underlying network which influences the computation times but also on the score distribution on the vertices of the network. Since citation networks have an intrinsic time arrow because of their immutable nature, the damping factor plays an important role in the score distribution over publication years. The larger the value of α, the more emphasis is given to early papers, whereas a value close to zero will remove the effect of the publication years since the ranking scores converge to $\frac{1}{n}$ for each paper.

Using the test data of award papers to evaluate the performance of PageRank with varying α values an optimal damping factor for a citation network can be obtained.

The graph in Figure 7.1(a) shows the precision and recall values of PageRank evaluated on the MAS CS data set. PageRank performs the best when $\alpha = 0.55$. With this damping value PageRank's precision is 42.51% and recall is 89.86% which is a slight improvement over using the default damping value. Furthermore, the AMAP@10 value is 0.61 when using PageRank with $\alpha = 0.55$ compared to 0.57 with $\alpha = 0.85$.

Using the DBLP citation network, PageRank performs the best with a damping factor of $\alpha = 0.25$ and achieves an AMAP@10 value of 0.65 by placing 90.07% of award papers in the top 10 ranks and 51.66% papers at position one. These results are displayed in Figure 7.1(b). Using $\alpha = 0.25$ only improves the precision but not the recall values.

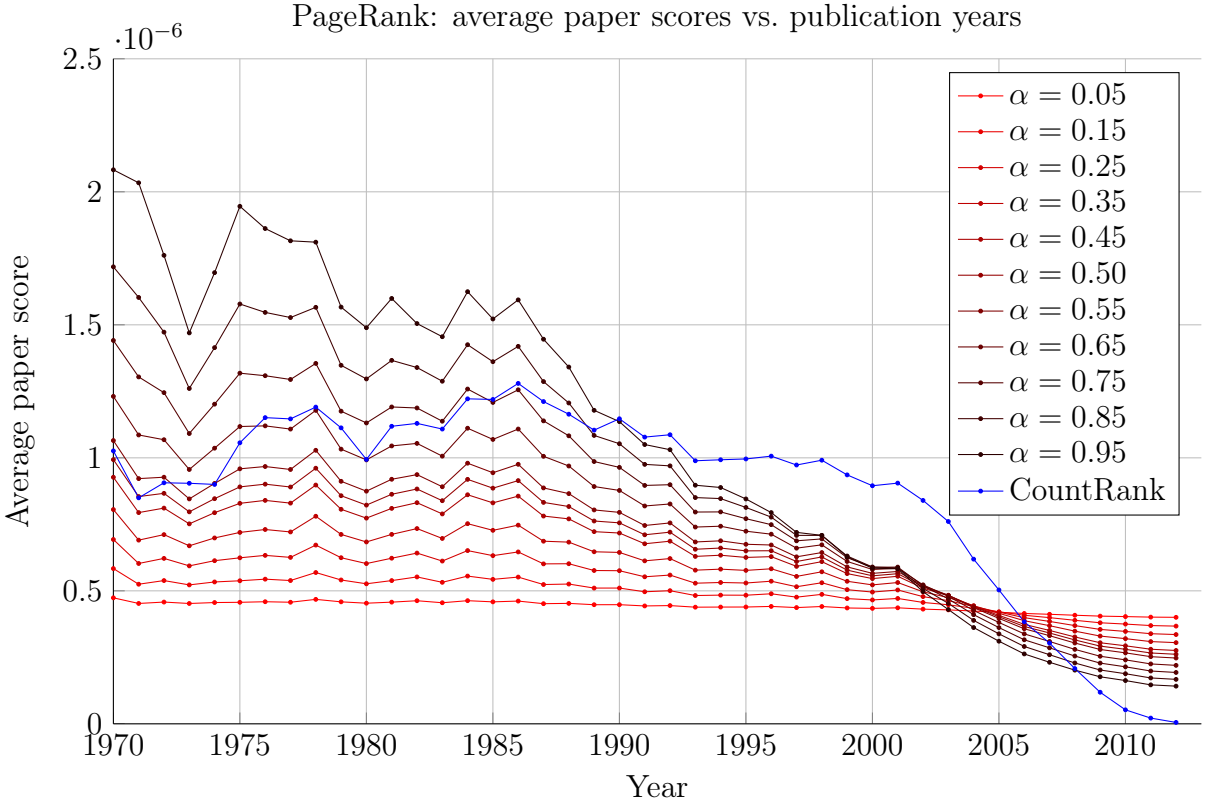


Figure 7.2: Average score distribution over publication years for PageRank with varying α values on the MAS CS citation network. For comparison the results of CountRank are included.

From Figure 7.2 one can see that for a small damping factor ($\alpha = 0.05$) the average score of papers over the years is relatively stable. Since the underlying citation network consists of 2 394 976 papers the average score if $\alpha \to 0$ is $\frac{1}{2\,394\,976} = 0.42 \cdot 10^{-6}$ which is in accordance with observed value in the figure.

Furthermore, in Figure 7.2 one can clearly see that the larger the damping factor is, the more weight is given to the scores of older papers. The highest Pearson correlation values between the average scores of CountRank and PageRank are obtained with $\alpha = 0.65$ ($r = 0.878$) and $\alpha = 0.75$ ($r = 0.880$), which would indicate that PageRank would perform the closest to CountRank for these α values. As mentioned earlier, PageRank performs the best with $\alpha = 0.55$ on the MAS citation network.

Lastly, the damping factor of PageRank also has an impact on the computation times as can be seen in Figure 7.3. The larger the value α, the more iterations are required by PageRank to achieve the same precision. For a damping factor of 0.95, PageRank needs many more iterations to obtain even a fairly low precision of 0.0001. This is due to the numerical instabilities introduced by the nature of the underlying citation network with such a damping factor since the network is very loosely connected.

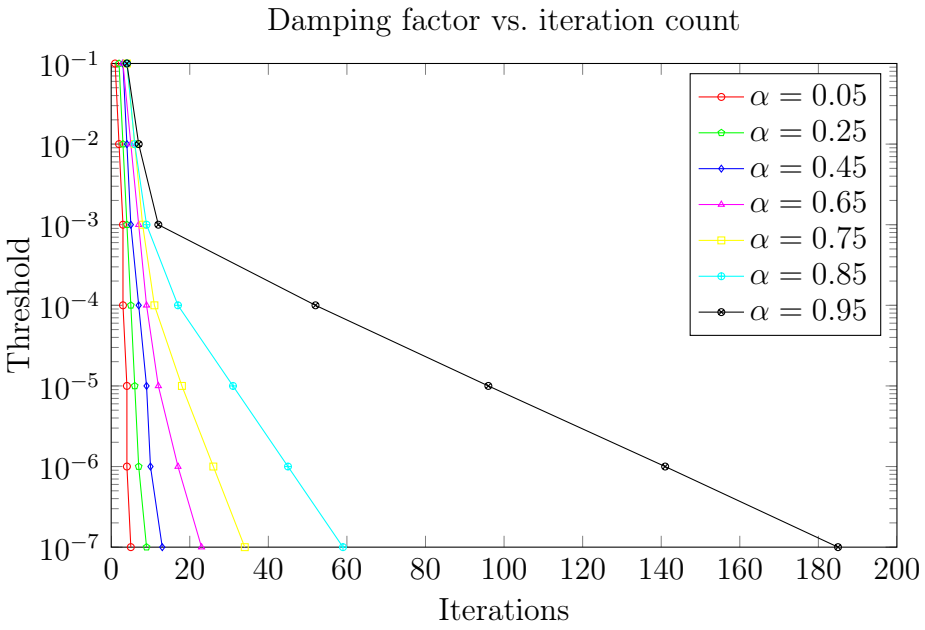


Figure 7.3: Number of iterations required by PageRank with varying damping factors to compute result values with a certain precision as indicated on the left hand side of the graph. PageRank was computed on the MAS CS network with order 2 394 976 and size 12 907 440 and 378 922 dangling vertices.

Optimising NewRank

In addition to the damping factor the NewRank algorithm has a second parameter τ controlling the decay time. The larger the value of τ, the smaller the influence of the age of a publication has on its ranking score. Moreover, if $\tau \to \infty$ then all references from a paper have the same weight and transfer an equal credit towards all cited papers. Since the damping factor α also has an effect on the score distribution over the citation network and is influenced by the publication dates of papers, they must be considered together.

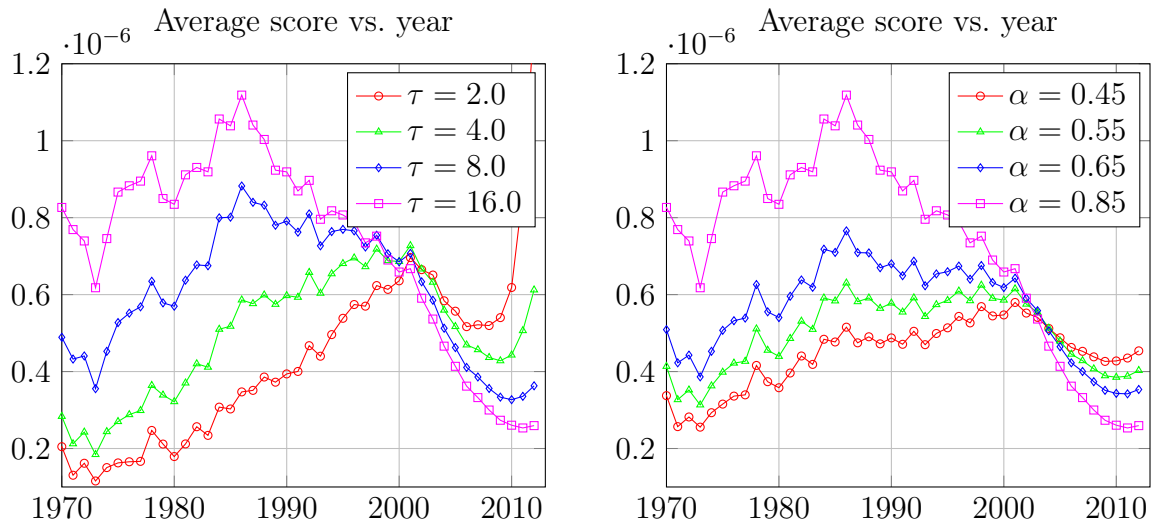
The graph in Figure 7.4(a) depicts the effect of varying τ for a fixed value of $\alpha = 0.85$. Similarly, the graph in Figure 7.4(b) depicts the change of the average score per year when α is varied and τ is set to 16.

Again using the award papers as evaluation data, NewRank performs best with a damping value of $\alpha = 0.55$ and τ set to 16. With these parameters NewRank obtains a precision of 42.51% and a recall value of 90.34%. The achieved AMAP@10 is 0.61 compared to 0.54 with the default values.

Summary of Optimising the Algorithms

Each bibliographic citation network possesses different properties. The article distribution over the publication years vary due to the age and prevalence of different academic fields. The same holds true for citation distributions of articles over time. In addition, citation practices differ between various disciplines. As a result, the ranking algorithms used on these citation networks have to be fine-tuned to fit the underlying citation network.

The best AMAP@10 value that PageRank and NewRank achieve on the MAS CS network is 0.61, which is obtained when both algorithms are initialised with a damping factor of $\alpha = 0.55$ and NewRank's time decay parameter is set to $\tau = 16$. Both algorithms



(a) Average score per publication year of papers using NewRank with a fixed damping value $\alpha = 0.85$ and varying τ values.

(b) Average score per publication year of papers using NewRank with varying damping values and a fixed time decay parameter of $\tau = 16.0$.

Figure 7.4: The effect of varying parameters of NewRank on the score distribution of papers over publication years.

Table 7.4: Summary of finding the optimal parameters for the algorithms on the complete MAS CS citation network.

Algorithm	Parameters	AMAP@10	F1 Score	Precision	Recall
CountRank	None	0.6146	0.5958	0.4444	0.9034
PageRank	$\alpha = 0.55$	0.6054	0.5772	0.4251	0.8986
NewRank	$\alpha = 0.45$, $\tau = 16.0$	0.6088	0.5782	0.4251	0.9034
YetRank	$\alpha = 0.25$, $\tau = 12.0$	0.6311	0.5881	0.4348	0.9082
SceasRank	$\alpha = 0.95$, $a = 2.5$, $b = 0$	0.6028	0.5870	0.4348	0.9034

obtain but do not improve on the precision and recall values of simply counting citations of papers.

SceasRank's optimal parameters for the MAS CS network are $\alpha = 0.95$ and $a = 2.5$. Using different variables for b had no effect on the ranking results of the award papers and therefore did not influence the evaluation metrics. Furthermore, no difference was observed between modifying the citation network by adding n edges to all dangling vertices and using an unmodified version. Note that NewRank achieves the same precision as PageRank and the same recall value as CountRank.

It should also be noted that optimal parameters for these algorithms are only valid for the specific underlying citation network. For citation networks with different properties such as the paper distribution over publication years and the citation distribution over years will have different optimal parameters.

7.2 How Well do Venues Predict High-Impact Papers?

Purpose

At many Computer Science conferences one or more papers are awarded a best paper prize in the year the papers are presented. Usually all papers presented in a year are considered for this award. Either a review panel of experts choose the best paper or the reviewers of the peer review processes give their recommendations on the quality of the papers to the conference panel from which the best papers are then chosen.

There are varying guidelines on how many best-paper awards are awarded. For example, at ICSE not more than 10% of papers are allowed to receive the prize. Alternatively, some conferences award a best paper award per track.

The set of papers that won the best paper awards is used to evaluate the precision of the conferences in predicting future high-impact papers.

Experimental Method

A list of 464 best papers from 32 conferences was collected and used as test data to evaluate how well the various conferences predict high-impact papers based on the CountRank algorithm since it performed the best in identifying high-impact papers (see Section 7.1). For each year that a conference awards best-paper prizes, the AP@10 of the best papers is calculated from the ranks of all papers published at the conference in that year. The MAP@10 is then calculated over all years in which best-paper awards were handed out.

Results

The results are organised into two groups of columns. In the left column, "Full Network", the precision values are given when used on the entire MAS CS citation network. The number of best papers in the test data for each conference is given in column "Count" and the average citation count that the best papers received is given in column "In-Degree".

Table 7.5: The precision of the award committees in identifying high-impact papers based on the papers that won best-paper awards at the associated conferences. The columns below "Full Network" displays the MAP values if the entire MAS CS citation network is used. Alternatively, "Subset Network" shows the precision values if the input network is truncated to 5 years after the papers won the best-paper awards. Only the top 5 venues are listed in this table. The entire listing of all 32 venues can be found in Appendix A.3.

	Full Network			Subset Network		
Conference	**Count**	**In-Degree**	**MAP@10**	**Count**	**In-Degree**	**MAP@10**
SOSP	22	118.36	0.595	19	66.89	0.577
OSDI	12	117.75	0.549	12	78.42	0.544
SIGMETRICS	8	67.75	0.486	7	54.71	0.525
FOCS	10	65.60	0.463	10	57.90	0.495
ACL	14	72.50	0.467	11	68.00	0.457

For the values given in the right column, "Subset Network", the MAS CS network is truncated to 5 years after the publication date of the papers that won the best-paper awards. This is done to see whether the publication dates of papers have an impact on the analysis since the year ranges for which best-paper awards were handed out are not

identical for each conference. For example, AAAI lists best papers since 1996 while for SIGMOBILE the data set only has best papers since 2008. See Table A.3 in Appendix A.2 for a detailed listing. This discrepancy is overcome by restricting the evaluation data to 5 years after publication. All best papers published after 2008 are ignored for the subset network since the MAS data only contains papers up to 2013.

From the table one can see that the venues SOSP (ACM Symposium on Operating Systems Principles), OSDI (Operating Systems Design and Implementation) and SIGMETRICS (Special Interest Group on Measurement and Evaluation) predict high-impact papers the most accurately with a MAP@10 of over 0.5 when using the subset network and also perform the best when using the entire MAS CS network.

Table 7.6: The precision of the top 5 award committees in identifying high-impact papers based on the single papers that won a best-paper award with the highest citation counts for each year in which the best-paper prize was awarded. The complete list of all 32 venues is given in Table A.11 in Appendix A.3.

Conference	Nr. Years	In-Degree	Avg. Papers	MAP@10
SOSP	7	106.50	2.71	0.640
SIGMETRICS	6	51.50	1.17	0.483
ACL	8	78.50	1.38	0.458
FOCS	6	68.33	1.67	0.410
FSE	7	71.57	2.43	0.405

On the one hand, it could be argued that the more papers that are awarded the best paper prize, the higher the chances of choosing papers that will not receive high citation counts. This can result in a lower precision. In order to account for this bias, only one best paper per year can be chosen for each conference. On the other hand, a venue that awards more best paper prizes in a year has a higher chance to choose the paper which receives the most citations in the following years.

One possible way of choosing a single best paper per year for each venue is by considering the papers' citation counts. For completeness, the results of choosing only one best paper with the highest citation count for each year are given in Table 7.6 which shows the top 5 venues that predicted the high-impact papers the most accurately.

The column "Nr. Years" shows for how many years the venues awarded best paper prizes. These values therefore indicate how many best papers are considered when computing the precision of how well the venues predict high-impact papers. Similarly to the previous result tables, the column "In-Degree" shows the average number of citations of the best papers that are chosen as test data. In this case, it shows the average citation count of the best papers with the most citations for each year at a venue. In the column "Avg. Papers", the average number of papers that are assigned best-paper prizes in a year are listed. Lastly, the column "MAP@10" shows the mean average precision of the venues in the prediction of high-impact papers.

One can see that the top conferences stay roughly the same. Again "SOSP" achieves the highest precision with 0.640. However, it should be noted that the precision values in Table 7.6 are notably higher than the values obtained in Table 7.5 where all best papers are considered. This is expected since only the papers with the highest citation counts are chosen for each year at each conference which are ranked higher than the other best papers that are ignored.

7.3 Evaluating Author Ranking Algorithms

Purpose

In order to objectively evaluate the ranking algorithms that rank venues, test data that contains information about the quality of journals or conferences is required. Since this type of data is not readily available, the ranking algorithms that rank authors are evaluated with appropriate test data. This is possible since the ranking algorithms for venues can also be adapted to rank authors since both entities publish one or more articles. The main difference is that authors can publish at different venues while journals and conferences intrinsically publish at a unique venue.

In this section the ranking algorithms that take author co-citation networks as input are evaluated against a list of authors that have won awards for their innovative, highly significant and enduring contributions to their fields of study.

Experimental Method

The test data is a set of authors that won author awards from at least one of the 19 different awards listed in Table A.4 in Appendix A.2. Of the in total 268 authors, 17 have won two different awards while "Karen Spärck Jones" won three awards, namely, the "ACM - AAAI Allen Newell Award", the "ACL Lifetime Achievement Award", and the "Gerard Salton Award" handed out by the "Special Interest Group on Information Retrieval" (SIGIR).

Since the authors that won the author awards are from various disciplines and the awards fall into different domains, all authors in the entire MAS CS citation network are considered when evaluating the author ranking algorithms. Therefore, the average and median ranks of the award authors are computed. Authors that won multiple awards are only counted once. Therefore, 249 authors are used for this evaluation.

Results

Table 7.7 lists the average and median ranks as produced by the various author ranking algorithms. The Author-Level Eigenfactor method achieves the best results with a median rank of 728.

Table 7.7: The results of evaluating the author ranking algorithms against the list of 249 authors that won innovation and contribution awards.

Algorithm	**Median Rank**
CCR	907
CCRS	925
AF	728
h-index	1 035
g-index	940
i10-index	1 371
PC	3 201

Using citation counts with self-citations omitted performs second best (907) followed by citation counts with author-self citations included. This indicates that self-citations do not necessarily increase an author's chance of receiving contribution awards. Further

investigation is required to measure the impact that author collaboration has on these results. The g-index ranks the award authors higher than the h-index. The worst indicator is using the publication counts of authors which is expected since the number articles that an author has published rather reflects the author's contribution to a field instead of the impact that the articles have on a field.

The AF metric ranks the award authors the highest with a median rank of 720 when the damping factor is set to 0.84.

7.4 How Well can Important Papers be Identified by Ranking Algorithms?

Purpose

Wikipedia lists the most important papers in different academic disciplines [68]. An important paper, according to the guidelines of Wikipedia, is a paper that led to new avenues of research, changed the scientific knowledge significantly, or had a substantial impact on the teaching of a domain. These papers were collected for Computer Science and matched against the paper entries in the MAS data set. A total of 129 of these papers are found in the MAS data set of which 115 papers contain year and venue values. This set of 115 papers is used to evaluate how well the various ranking algorithms can identify important papers.

Experimental Method

Since the set of important papers span various fields in Computer Science and are published in different journals and conferences, the overall ranks of the papers are used as a metric to evaluate the ranking algorithms independent of the publication years of the papers. Therefore, the median rank of the important papers is computed on the whole set of 1 573 679 CS papers. It should be noted that the average publication year of the important papers is 1981 which is relatively old.

Results

Using the default parameter values for the algorithms, PageRank ranks the important papers the highest with a median rank of 990, followed by YetRank (1078) and CountRank (1652). NewRank performs the worst (9566) which can be explained by the fact that the average publication year of the important papers is 1981 and NewRank gives higher priority to recently published papers.

Table 7.8 shows the results of the algorithms' default and optimal parameters for identifying the important papers. For both YetRank and NewRank the optimal α and τ values are large with ($\alpha = 0.85$, $\tau = 40$) and ($\alpha = 0.95$, $\tau = 60$), respectively. Since the important papers are relatively old, these values are expected since they shift the focus towards older publications in the citation network as shown in Figures 7.2 and 7.4.

It should be noted that for PageRank the average rank of the important papers decreases the larger α becomes but that the smallest median rank is found with $\alpha = 0.85$. The same trend can be observed with NewRank. In addition, keeping α the same value, the median rank decreases by choosing larger τ values. By increasing the τ values, the effect that the age of a publication has on the resulting scores of papers is decreased. This

Table 7.8: Results of evaluating the ranking algorithms against a set of 115 important papers in Computer Science. The median rank of the important papers is used to measure the algorithms' precision. The results of using the default values are listed on the left. On the right, the parameter values for which the algorithms ranked the important papers the highest are given.

Algorithm	Default Parameters	Median	Optimal Parameters	Median
CountRank	None	1652	–	–
PageRank	$\alpha = 0.85$	990	$\alpha = 0.85$	990
NewRank	$\alpha = 0.85$, $\tau = 4.0$	9566	$\alpha = 0.95$, $\tau = 60.0$	1179
YetRank	$\alpha = 0.85$, $\tau = 4.0$	1078	$\alpha = 0.85$, $\tau = 40.0$	807
SceasRank	$\alpha = 0.85$, $a = e$, $b = 1.0$	2153	$\alpha = 1.0$, $a = 1.05$, $b = 0.5$	1080
SceasRank1	$\alpha = 1.0$, $a = e$, $b = 1.0$	1898	–	–
SceasRank2	$\alpha = 0.85$, $a = e$, $b = 0$	2153	–	–

indicates that for this set of important papers, the age of publications is not as important as the citations they receive.

SceasRank performs the best with $a = 1.05$ and $\alpha = 1$. As seen in previous experiments the value of b has no effect on the results as long as $b > 0$. Similarly, adding outgoing edges to the dangling vertices in the network also does not impact the results when using SceasRank.

All algorithms perform better than CountRank after finding optimal parameters for each. The overall best performing algorithm is YetRank which ranks the important papers at a median rank of 807.

7.5 Chapter Summary

This chapter covered various experiments that evaluate the ranking algorithms using different test data sets. Different algorithms were found to be best suited for different experiments. The following list summarises the results found in this chapter:

- CountRank performs the best in identifying high-impact award papers when used on the full MAS CS and DBLP citation networks. On the MAS and DBLP citation networks CountRank achieves MAP values of 0.61 and 0.66, respectively.
- When the MAS CS citation network is truncated to the years that the high-impact paper prizes are awarded, YetRank outperforms CountsRank. YetRanks obtains a MAP of 0.61 while CountRank achieves a MAP value of 0.59.
- After finding the optimal parameters for all algorithms to identify the set of high-impact papers, YetRank achieves the highest MAP with 0.63.
- The SOSP (ACM Symposium on Operating Systems Principles) conference predicts high-impact papers the most accurately with a MAP value of 0.64.
- The Author-Level Eigenfactor is the best metric at identifying the set of authors that won innovation and contribution awards. The best result was achieved with a damping factor of $\alpha = 0.84$ where the award authors are ranked with a median rank of 720.

- Using the algorithms' default parameter values, PageRank identifies the overall important papers in Computer Science the most accurately with a median rank of 990.
- After finding the optimal parameters for all algorithms to find the overall important papers, YetRank achieves the best median rank of 807.
- It should be noted that the performance of the PageRank-like algorithms are very sensitive to the parameters that are used. The algorithms that incorporate a time decay parameter, namely NewRank and PageRank, are effected even more by the sensitivity of the parameters.

Chapter 8

Conclusion

In this study we presented an in-depth discussion on what citation counts can measure and what aspects influence the citation counts of papers, authors and venues. Furthermore, popular bibliometric measures and recently proposed ranking algorithms were discussed, categorised and formulated mathematically.

We compared these ranking algorithms on a Computer Science citation network obtained from Microsoft Academic Search and identified various ranking properties of the algorithms. Furthermore, we presented results from evaluating paper and author ranking algorithms against test data sets that are based on expert opinions.

8.1 Summary of Findings

By evaluating the author ranking algorithms with a set of authors that won contribution awards, we found that the Author-Level Eigenfactor metric (AF) identifies these authors the most accurately. The optimal α value for AF is 0.84 which is very close to the default value of 0.85. When constructing an author co-citation graph from citation data, the intrinsic time-arrow exhibited by paper citation networks falls away. Therefore, it is not surprising that the optimal α value is close to 0.85 which is also the default value of PageRank initially used by Google for the Internet's hyperlink graph [4].

We found that the damping factor α of PageRank-like algorithms also plays an important role when used on bibliographic citation networks to rank papers. The choice of α impacts the results heavily and is influenced by the publication dates of papers and the structure of the underlying citation network.

When considering PageRank, for example, different optimal α values were found for different purposes. The optimal damping factor for identifying current research activities is 0.25. For finding papers that won high-impact prizes the best value is 0.55 and for identifying overall important papers it is 0.85.

Empirically, these parameter values are consistent with the observation that α controls the score distribution over the years. The larger the value of α, the higher the scores of older papers. The current research activity is immediate where recently published papers are more relevant and therefore the optimal value for α is relatively small. Papers receive high-impact prizes about 10 to 15 years after their publication and fall within the mid-range of all published papers in the data set. Accordingly, the optimal α value was found to be 0.55. Lastly, the set of important papers are relatively old, with an average publication year of 1981, and hence the optimal value of 0.85 is comparatively large.

The optimal α parameter for PageRank-like algorithms also changes when used on different citation networks. Using the MAS CS citation network, PageRank identifies the award papers most accurately, with a damping factor of 0.55. Alternatively, when using the DBLP data set, the optimal α value was found to be 0.25 for the same experiment.

8.2 Threats to Validity

For all the experiments in Chapters 6 and 7, the CS subset of the MAS data set was used. Therefore, only citations are used that originate from CS papers or are citations that directly cite CS papers. This means that all citations that originate from outside the CS domain are weighted the same, which does not reflect the true weight if the entire citation network would have been considered. Therefore, using the CS citation network has to be seen as an approximation of the entire academic citation network structure. Because of the time and space complexity of the algorithms it was not feasible to compute the various ranking algorithms on the entire citation network. Furthermore, the validity of the results discussed in this paper is dependent on the data quality of the citation databases used.

The use of award papers that won prizes retrospectively for their high impact is not perfect test data. For most venues the selection process requires someone to submit potential papers manually to the review panel. The selection of the final award papers is therefore subject to the the submission process. High-impact papers might not be considered since they were not submitted for evaluation in the first place.

The set of author awards used as test data are awarded to authors for their long-lasting, significant and innovative contributions to their field of study. This is also not perfect evaluation data. The selection of award authors is very subjective and takes other aspects of impact into consideration, in addition to the objective measures such as publication counts or the intrinsic quality of an author's work. For example, teaching duties and administrative work are also considered as contributions of a researcher and cannot be measured based on his or her publication record. Furthermore, all author awards are treated equally but some prizes might be more prestigious than others.

8.3 Contributions

The following list summarises the contributions that emerged from the research presented in this thesis:

- This thesis provides an in-depth discussion on the aspects that influence the citation counts of papers. Furthermore, a review on what citation counts can and cannot measure is given.
- Bibliometric measures and ranking algorithms that rank academic entities such as papers, authors and journals are categorised and defined in a concise and uniform mathematical formulation. The list of metrics included in this paper is not exhaustive. Only better known and often used metrics that are based on pure citation counts are included. The focus is on ranking algorithms that are based on PageRank that use bibliometric citation networks to model the scientific processes.
- The Microsoft Academic Search data set is used to analyse a variety of publication trends. The papers in the data set are categorised into disciplines to show differences

in the production of scientific output between these disciplines. Moreover, trends that show how research has changed in the last decades are presented.

- The collection of a large test data set of important papers and influential authors that can be used for research dealing with ranking algorithms on academic citation networks.
- The above-mentioned test data sets are used to evaluate paper and author ranking algorithms discussed in this paper. The best suited algorithms for different applications are identified.

8.4 Suggestions for Future Work

The MAS data set is in the public domain for research purposes since the middle of 2014. It is presumably the largest free citation database available and will spur a lot of new research in bibliometrics and scientometrics. The challenge is to use the data set appropriately. Citation data is not perfect and in order to obtain valid and exact results, the MAS data has to be cleaned before it can be used for citation analysis.

As mentioned in Section 5.4, the MAS dataset exhibits an anomaly in the citation data that persists despite our efforts to identify the source. The anomaly occurs between 1994 and 1995 when any measurements are plotted against years. For example, when the number of publications produced in a year are plotted, a sudden increase can be observed in 1995. It was found that this sudden increase is independent of the field of study and is exhibited by papers from most publishing sources. However, the increase is more severe depending on the publishing sources and the discipline.

Curiously, the anomaly is also observed if trends are plotted against the time since an authors first publication and not against the calendar years. Here it was observed that a sudden increase occurs exactly 20 years after an author's first publication.

We assume that this anomaly stems from the internal indexing of the MAS data and not from a single publishing source. We base this assumption on the following empirical observations:

- The anomaly is exhibited the least by papers from Social Science, Economics & Business, and Arts & Humanities.
- The largest jump is observed by Biology, Chemistry, Engineering and Physics papers.
- The disciplines that exhibit this anomaly heavily obtain most of their papers from Elsevier and Springer (on average 75.31%).
- However, Elsevier and Springer also contribute a large part of papers to the disciplines that exhibit the anomaly the least (on average 24.53%).
- The Computer Science domain exhibits the anomaly as well and 80.37% of their papers are sourced from DBLP and only 7.08% from Elsevier and Springer.
- All other publishing sources are too small or domain-specific to be the cause for this anomaly.

For future research using the MAS data set, it is important to identify the cause of this anomaly in order to obtain more accurate results. It is also possible to find additional systematic data errors in the MAS data. Therefore, a standardised oracle test data set is required to quantify the quality of the data set. A representative number of papers has to be sampled from the data set and their citation, year and author information validated manually. This sample data can be used as test data to improve the data preprocessing such as author-name disambiguation and paper title merging.

Another problem is that the MAS data is not updated and does not contain new papers published after 2013. Therefore, current research trends cannot be conducted on the MAS data and it has to be seen as historic data.

Additional and more in-depth trend analysis between different disciplines on the MAS data is an interesting research avenue on its own. Moreover, since the publication patterns and citation conventions differ between various academic disciplines, the effects of these aspects have to be identified, analysed and incorporated into the computations of new and better ranking algorithms.

The test data that was collected for the purpose of this thesis is not exhaustive. Additional test data can be obtained by retrieving papers and authors that won awards from other conferences and organisations. Furthermore, the test data can be extended to span across domains and not only Computer Science data. This becomes especially important if ranking algorithms are computed over multiple disciplines.

The MAS data set classifies papers into top-level disciplines such as Computer Science and Mathematics. These domains are rigid and a paper's categorisation depends on the venue at which it was published. Papers and authors can also be clustered programmatically using techniques proposed by [71; 72; 73] to reveal more fine-grained clusters of topics and community structures.

The challenge after clustering papers into their topics is the programmatic labeling of the clusters without the need for human evaluation. This could be achieved by using term counts of the words in the papers' titles and abstracts. This labeling technique can be used in combination with the keywords associated with the papers. Finding appropriate solutions for this task also suggests interesting research avenues for the future.

Bibliography

[1] Gross, P.L.K. and Gross, E.M.: College libraries and chemical education. *Science*, vol. 66, no. 1713, pp. 385–389, 1927.

[2] Garfield, E.: Citation indexes for science. A new dimension in documentation through association of ideas. *Nature*, vol. 122, no. 3159, pp. 108–111, 1955.

[3] Hirsch, J.: An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 46, pp. 16569–16572, 2005.

[4] Brin, S. and Page, L.: The anatomy of a large-scale hypertextual web search engine. In: *Proceedings of the Seventh International Conference on World Wide Web*, WWW '07, pp. 107–117. Elsevier Science Publishers B. V., Amsterdam, The Netherlands, 1998.

[5] Microsoft Research: Microsoft Academic Search. `http://academic.research.microsoft.com`, 2013. [Online; accessed 16-May-2014].

[6] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L. and Su, Z.: Arnetminer: Extraction and mining of academic social networks. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pp. 990–998. ACM, New York, USA, 2008.

[7] The DBLP Team: The DBLP Computer Science Bibliography. `http://dblp.uni-trier.de/`, 2014. [Online; accessed 16-May-2014].

[8] Garfield, E.: Is citation analysis a legitimate evaluation tool? *Scientometrics*, vol. 1, no. 4, pp. 359–375, 1979.

[9] Bergstrom, C.T., West, J.D. and Wiseman, M.: The eigenfactor metrics. *The Journal of Neuroscience*, vol. 28, no. 45, pp. 11433–11434, 2008.

[10] West, J.D., Jensen, M.C., Dandrea, R.J., Gordon, G.J. and Bergstrom, C.T.: Author-level eigenfactor metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community. *Journal of the American Society for Information Science and Technology*, vol. 64, no. 4, pp. 787–801, 2013.

[11] Sidiropoulos, A. and Manolopoulos, Y.: A citation-based system to assist prize awarding. *SIGMOD Records*, vol. 34, no. 4, pp. 54–60, 2005.

[12] Chen, P., Xie, H., Maslov, S. and Redner, S.: Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, vol. 1, no. 1, pp. 8–15, 2007.

[13] Walker, D., Xie, H., Yan, K.-K. and Maslov, S.: Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, vol. 6, no. P06010, 2007.

[14] Hwang, W., Chae, S., Kim, S. and Woo, G.: Yet another paper ranking algorithm advocating recent publications. In: *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 1117–1118. ACM, New York, USA, 2010.

[15] Dunaiski, M. and Visser, W.: Comparing paper ranking algorithms. In: *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference*, SAICSIT '12, pp. 21–30. ACM, New York, USA, 2012.

[16] Martin, B.R.: The use of multiple indicators in the assessment of basic research. *Scientometrics*, vol. 36, no. 3, pp. 343–362, 1996.

[17] Hood, W.W. and Wilson, C.S.: The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*, vol. 52, no. 2, pp. 291–314, 2001.

[18] Garfield, E.: From the science of science to scientometrics visualizing the history of science with histcite software. *Journal of Informetrics*, vol. 3, no. 3, pp. 173 – 179, 2009.

[19] Tague-Sutcliffe, J.: An introduction to informetrics. *Information Processing & Management*, vol. 28, no. 1, pp. 1–3, 1992.

[20] Wilson, C.S.: Informetrics. *Annual Review of Information Science and Technology*, vol. 34, pp. 107–247, 2001.

[21] Kessler, M.: Bibliographic coupling between scientific papers. *American Documentation*, vol. 14, no. 1, pp. 10–25, 1963.

[22] Small, H.: Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, vol. 24, no. 4, pp. 265–269, 1973.

[23] Winston, W.L.: *Operations Research: Applications and Algorithms, Fourth Edition*. Brooks/Cole, Bellmont, CA, USA, 2004.

[24] Lee, C.P., Golub, G.H. and Zenios, S.A.: A fast two-stage algorithm for computing PageRank and its extensions. Tech. Rep. SCCM-2003-15, Scientific Computation and Computational Mathematics, Stanford University, 2003.

[25] Ipsen, I.C.F. and Kirkland, S.: Convergence analysis of a pagerank updating algorithm by Langville and Meyer. *SIAM Journal on Matrix Analysis and Applications*, vol. 27, no. 4, pp. 952–967, 2006.

[26] Langville, A. and Meyer, C.: Deeper inside PageRank. *Internet Mathematics*, vol. 1, no. 3, pp. 335–380, 2004.

[27] Brodman, E.: Choosing physiology journals. *Bulletin of the Medical Library Association*, vol. 32, no. 4, pp. 479–483, 1944.

[28] Lawani, S.M.: On the heterogeneity and classification of author self-citations. *Journal of the American Society for Information Science*, vol. 33, no. 5, pp. 281–284, 1982.

[29] Phelan, T.: A compendium of issues for citation analysis. *Scientometrics*, vol. 45, no. 1, pp. 117–136, 1999.

[30] Aksnes, D.W.: A macro study of self-citation. *Scientometrics*, vol. 56, no. 2, pp. 235–246, 2003.

[31] Suber, P.: Gratis and libre open access. `http://www.sparc.arl.org/resource/gratis-and-libre-open-access`, 2008. [Online; accessed 20-Aug-2014].

[32] Suber, P.: Open Access Overview. `http://legacy.earlham.edu/~peters/fos/overview.htm`, 2013. [Online; accessed 1-Jul-2014].

[33] Lawrence, S.: Free online availability substantially increases a paper's impact. *Nature*, vol. 411, no. 6837, p. 521, 2001.

[34] Brody, T. and Harnad, S.: Comparing the impact of open access (OA) vs. non-OA articles in the same journals. *D-Lib Magazine*, vol. 10, no. 6, 2004.

[35] Cornell University Library: arXiv Primer. `http://arxiv.org/help/primer`, 2014. [Online; accessed 1-Jul-2014].

[36] Kurtz, M.: Restrictive access policies cut readership of electronic research journal articles by a factor of two. `http://opcit.eprints.org/feb19oa/kurtz.pdf`, 2004. [Online; accessed 20-Aug-2014].

[37] Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E. and Murray, S.S.: The effect of use and access on citations. *Information Processing and Management*, vol. 41, no. 6, pp. 1395–1402, 2005.

[38] Hajjem, C., Harnad, S. and Gingras, Y.: Ten-year cross-disciplinary comparison of the growth of open access and how it increases research citation impact. *arXiv preprint cs/0606079*, 2006.

[39] Moed, H.F.: The effect of "open access" on citation impact: An analysis of ArXiv's condensed matter section. *Journal of the American Society for Information Science and Technology*, vol. 58, no. 13, pp. 2047–2054, 2007.

[40] Davis, P.M.: Open access, readership, citations: a randomized controlled trial of scientific journal publishing. *FASEB Journal*, vol. 25, no. 7, pp. 2129–34, 2011.

[41] Falagas, M.E., Pitsouni, E.I., Malietzis, G.a. and Pappas, G.: Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *FASEB Journal*, vol. 22, no. 2, pp. 338–42, 2008.

[42] Zhang, L.: The impact of data source on the ranking of computer scientists based on citation indicators: A comparison of Web of Science and Scopus. *Science & Technology Librarianship*, 2014.

[43] Elsevier: Scopus - Content Overview. `http://www.elsevier.com/online-tools/scopus/content-overview`, 2014. [Online; accessed 17-Jul-2014].

[44] Thomson Reuters: Web Of Science - Core Collection. `http://thomsonreuters.com/products/ip-science/04_064/wos-core-collection.pdf`, 2014. [Online; accessed 17-Jul-2014].

[45] Franceschet, M.: A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. *Scientometrics*, vol. 83, no. 1, pp. 243–258, 2010.

[46] Kulkarni, A., Aziz, B., Shams, I. and Busse, J.: Comparisons of citations in Web of Science, Scopus, and Google Scholar for articles published in general medical journals. *JAMA: The Journal of the American Medical Association*, vol. 302, no. 10, 2009.

[47] The College of Information Sciences and Technology: CiteSeer$^x\beta$. `http://citeseerx.ist.psu.edu`, 2013. [Online; accessed 16-May-2014].

[48] Google Scholar: Google Scholar. `http://scholar.google.co.za/`, 2014. [Online; accessed 22-July-2013].

[49] Very Large Data Base Endowment Inc.: VLDB 10 years awards. `http://vldb.org/archives/10year.html`, 2014. [Online; accessed 16-May-2014].

[50] ACM Special Interest Group on Management of Data: SIGMOD awards. `http://www.sigmod.org/sigmod-awards/`, 2014. [Online; accessed 28-May-2014].

[51] Garfield, E.: The Agony and the Ecstasy - The History and Meaning of the Journal Impact Factor. `http://garfield.library.upenn.edu/papers/jifchicago2005.pdf`, 2005. [Online; accessed 20-Aug-2014].

[52] Garfield, E.: The Thomson Reuters Impact Factor. `http://wokinfo.com/essays/impact-factor/`, 1994. [Online; accessed 22-July-2013].

[53] Bollen, J., Rodriquez, M.A. and van de Sompel, H.: Journal status. *Scientometrics*, vol. 69, no. 3, pp. 669–687, 2006.

[54] Althouse, B.M., West, J.D., Bergstrom, C.T. and Bergstrom, T.: Differences in impact factor across fields and over time. *Journal of the American Society for Information Science and Technology*, vol. 60, no. 1, pp. 27–34, 2009.

[55] Van Nierop, E.: Why do statistics journals have low impact factors? *Statistica Neerlandica*, vol. 63, no. 1, pp. 52–62, 2009.

[56] Thomson Reuters: Journal citation reports. `http://thomsonreuters.com/journal-citation-reports`, 2014. [Online; accessed 16-May-2014].

[57] Bergstrom, C.: Eigenfactor: Measuring the value and prestige of scholarly journals. *College & Research Libraries News*, vol. 68, no. 5, pp. 314–316, 2007.

[58] Connor, J.: Google Scholar Citations Open To All. `http://googlescholar.blogspot.com/2011/11/google-scholar-citations-open-to-all.html`, 2011. [Online; accessed 16-May-2014].

[59] Meho, L.I. and Rogers, Y.: Citation counting, citation ranking, and h-index of human-computer interaction researchers: A comparison of Scopus and Web of Science. *Journal of the American Society for Information Science and Technology*, vol. 59, no. 11, pp. 1711–1726, 2008.

[60] Egghe, L.: Theory and practise of the g-index. *Scientometrics*, vol. 69, no. 1, pp. 131–152, 2006.

[61] Maslov, S. and Redner, S.: Promise and pitfalls of extending Google's PageRank algorithm to citation networks. *The Journal of Neuroscience*, vol. 28, no. 44, pp. 11103–11105, 2008.

[62] Xie, H., Yan, K.-K. and Maslov, S.: Optimal ranking in networks with community structure. *Physica A: Statistical Mechanics and its Applications*, vol. 373, pp. 831–836, 2007.

[63] de Solla Price, D.: Networks of scientific papers. *Science*, vol. 149, no. 3683, pp. 510–515, 1965.

[64] Redner, S.: Citation statistics from more than a century of physical review. *arXiv preprint physics/0407137*, 2004.

[65] Sidiropoulos, A. and Manolopoulos, Y.: Generalized comparison of graph-based ranking algorithms for publications and authors. *Journal of Systems and Software*, vol. 79, no. 12, pp. 1679–1700, 2006. ISSN 0164-1212.

[66] Bergstrom, C.T. and West, J.D.: Eigenfactor score and article influence score: Detailed methods. `http://www.eigenfactor.org/methods.pdf`, 2008. [Online; accessed 21-August-2013].

[67] Association for Computing Machinery: ACM Digital Library. `http://dl.acm.org/`, 2014. [Online; accessed 28-Aug-2014].

[68] Wikipedia: Lists of important publications in science. `http://en.wikipedia.org/wiki/Lists_of_important_publications_in_science`, 2014. [Online; accessed 16-May-2014].

[69] Association for Computing Machinery: Alphabetical listing of A.M. Turing award winners. `http://amturing.acm.org/alphabetical.cfm`, 2014. [Online; accessed 11-Aug-2014].

[70] Tamura, K., Stecher, G., Peterson, D. and Kumar, S.: MEGA: molecular evolutionary genetics anlysis. `http://www.megasoftware.net`, 2014. [Online; accessed 11-Aug-2014].

[71] Blondel, V.D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E.: Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[72] Zhang, L., Janssens, F., Liang, L. and Glänzel, W.: Journal cross-citation analysis for validation and improvement of journal-based subject classification in bibliometric research. *Scientometrics*, vol. 82, no. 3, pp. 687–706, 2010.

[73] Rosvall, M. and Bergstrom, C.T.: Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS One*, vol. 6, no. 4, p. e18209, 2011.

[74] Association for the Advancement of Artificial Intelligence: AAAI classic paper award. `http://www.aaai.org/Awards/classic.php`, 2014. [Online; accessed 16-May-2014].

[75] Striewe, M.: Most influential papers of automated software engineering. `http://ase-conferences.org/Mip.html`, 2014. [Online; accessed 16-May-2014].

[76] Special Interest Group on Programming Languages: Most influential ICFP paper award. `http://www.sigplan.org/Awards/Conferences/ICFP/Main`, 2014. [Online; accessed 16-May-2014].

[77] ACM Special Interest Group on Software Engineering: ICSE's most influential paper award. `http://www.sigsoft.org/awards/mostInfPapAwd.htm`, 2014. [Online; accessed 16-May-2014].

[78] SIGARCH: ACM SIGARCH/IEEE-CS TCCA influential ISCA paper award. `http://www.sigarch.org/awards/acm-sigarchieee-cs-tcca-influential-isca-paper-award/`, 2014. [Online; accessed 16-May-2014].

[79] Special Interest Group on Programming Languages: Most influential OOPSLA paper award. `http://www.sigplan.org/Awards/Conferences/OOPSLA/Main`, 2014. [Online; accessed 16-May-2014].

[80] Special Interest Group on Programming Languages: Most influential PLDI paper award. `http://www.sigplan.org/Awards/Conferences/PLDI/Main`, 2014. [Online; accessed 16-May-2014].

[81] Special Interest Group on Programming Languages: Most influential POPL paper award. `http://www.sigplan.org/Awards/Conferences/POPL/Main`, 2014. [Online; accessed 16-May-2014].

[82] ACM Special Interest Group for Genetic and Evolutionary Computation: SIGEVO impact award. `http://www.sigevo.org/wiki/tiki-index.php?page=SIGEVO+Impact+Award`, 2014. [Online; accessed 16-May-2014].

[83] ACM SIGCOMM: ACM SIGCOMM test of time paper award. `http://www.sigcomm.org/awards/test-of-time-paper-award`, 2014. [Online; accessed 16-May-2014].

[84] ACM SIGMETRICS: The SIGMETRICS test of time award. `http://www.sigmetrics.org/awards.shtml#testoftime`, 2014. [Online; accessed 16-May-2014].

[85] ACM Special Interest Group on Software Engineering: ACM SIGSOFT impact paper award. `http://www.sigsoft.org/awards/ImpactAward.htm`, 2014. [Online; accessed 16-May-2014].

[86] Huang, J.: Best Paper Awards in Computer Science. `http://jeffhuang.com/best_paper_awards.html`, 2014. [Online; accessed 20-Aug-2014].

[87] Striewe, M.: Automated software engineering conference. `http://www.ase-conferences.org/olbib/index.html`, 2014. [Online; accessed 28-May-2014].

[88] SIGSOFT: Winners of the ACM SIGSOFT distinguished paper award. `http://www.sigsoft.org/awards/disPapAwd-rec.htm`, 2014. [Online; accessed 28-May-2014].

[89] IEEE Computer Society: IEEE data mining awards. `http://www.cs.uvm.edu/~icdm/Awards.shtml`, 2014. [Online; accessed 28-May-2014].

[90] USENIX - The Advanced Computing Systems Association: Best papers. `https://www.usenix.org/conferences/best-papers`, 2014. [Online; accessed 28-May-2014].

[91] ACM Special Interest Group on Mobility of Systems, Users, Data, and Computing: MobiCom best paper award. `http://www.sigmobile.org/awards/mobicombestpaper.html`, 2014. [Online; accessed 28-May-2014].

[92] Association for Computing Machinery: Awards - ACM - AAAI Allen Newell award. `http://awards.acm.org/newell/year.cfm`, 2014. [Online; accessed 28-May-2014].

[93] The Association for Computational Linguistics: ACL lifetime achievement award recipients. `http://aclweb.org/aclwiki/index.php?title=ACL_Lifetime_Achievement_Award_Recipients`, 2013. [Online; accessed 28-May-2014].

[94] ACM Special Interest Group on Computer-Human Interaction: SIGCHI awards. `http://www.sigchi.org/about/awards`, 2014. [Online; accessed 28-May-2014].

[95] IEEE Computer Society: PAMI Azriel Rosenfeld lifetime achievement award. `http://www.computer.org/portal/web/tcpami/Azriel-Rosenfeld-Life-Time-Achievement-Award`, 2014. [Online; accessed 28-May-2014].

[96] International Joint Conferences on Artificial Intelligence: IJCAI-13 award for research excellence. http://www.ijcai.org/awards/, 2014. [Online; accessed 28-May-2014].

[97] SIGARCH: ACM SIGARCH Maurice Wilkes award past winners. http://www.sigarch.org/awards/acm-sigarch-maurice-wilkes-award-past-winners/, 2014. [Online; accessed 28-May-2014].

[98] SIGKDD: SIGKDD innovation award. http://www.kdd.org/sigkdd-innovation-award, 2014. [Online; accessed 28-May-2014].

[99] SIGPLAN - Special Interest Group on Programming Languages: Programming languages achievement award. http://www.sigplan.org/awards/achievement/main, 2014. [Online; accessed 28-May-2014].

[100] Chakrabarti, A.: Knuth prize. http://www.sigact.org/Prizes/Knuth/, 2014. [Online; accessed 28-May-2014].

[101] SIGCOMM: SIGCOMM award recipients. http://www.sigcomm.org/awards/sigcomm-awards, 2014. [Online; accessed 28-May-2014].

[102] SIGIR - Special Interest Group on Information Retrieval: Awards. http://sigir.org/general-information/awards/, 2014. [Online; accessed 28-May-2014].

[103] Wikipedia: SIGMETRICS. http://en.wikipedia.org/wiki/SIGMETRICS#Achievement_award, 2014. [Online; accessed 28-May-2014].

[104] ACM Special Interest Group on Mobility of Systems, Users, Data, and Computing: Outstanding contribution award. http://www.sigmobile.org/awards/oca.html, 2014. [Online; accessed 28-May-2014].

[105] Association for Computing Machinery, Inc.: ACM SIGSIM awards. http://www.acm-sigsim-mskr.org/awards.htm, 2014. [Online; accessed 28-May-2014].

[106] ACM Special Interest Group on Software Engineering: ACM SIGSOFT outstanding research award. http://www.sigsoft.org/awards/outResAwd.htm, 2014. [Online; accessed 28-May-2014].

[107] USENIX Association: Flame award. https://www.usenix.org/about/flame, 2014. [Online; accessed 28-May-2014].

Appendices

Appendix A

Additional Information and Results

In the following section information about the MAS data set is listed that is not shown in the main body of this thesis. In Section A.2 details about the test data sets are given, such as the name of the awards that were chosen and the information source from where the data was obtained. Additional results of the experiments conducted in this thesis are given in Section A.3. For example, the 10 highest ranked papers according to the paper rankings algorithms are listed. In addition, the top 10 authors according to the Author-Level Eigenfactor metric are given with the corresponding ranks that the authors receive according to citation counts, the h-index, the g-index and the $i10$-index.

A.1 Additional MAS Data Set Information

Table A.1 lists additional properties about the MAS data set categorised into the different domains.

Table A.1: Detailed information on the sizes of the different domain in the MAS data set.

Domain	Papers	Citations	Authors	Conferences	Journals
Agriculture Science	454,884	2,535,416	431,651	0	402
Arts & Humanities	1,349,265	1,181,863	535,137	0	1864
Biology	3,649,664	37,829,536	2,935,402	2	2272
Chemistry	4171802	20,453,946	2,826,268	0	856
Computer Science	2,245,128	21,511,117	1,152,558	3152	1351
Economics & Business	817,709	6,853,791	443,853	1	1425
Engineering	2,044,780	7,445,687	1,794,716	599	1447
Environmental Sciences	422,857	3,389,313	486,511	0	360
Geosciences	740,875	6,450,496	523,999	1	518
Material Science	842,997	2,530,727	778,286	0	363
Mathematics	989,423	5,410,431	371,548	1	627
Medicine	11,097,095	80,698,411	5,956,044	7	5765
Physics	2,001,157	6,214,444	1,346,109	1	725
Social Science	1,725,171	6,204,152	1,004,578	0	2242

The column "Papers" shows the number of papers that are associated with a publication date and are included into the data sets used for the experiments in this thesis. The citation counts displayed in the third column are citations from papers that cite a paper in the Computer Science domain. Therefore, if a Computer Science paper references a

paper that falls outside of the Computer Science domain, the corresponding citation is not included in this citation count.

It should be noted that the paper counts displayed in Table A.1 do not reflect the sizes of the various disciplines. Firstly, the papers are categorized into domains depending on the venue at which they are published. Generally this is a good categorisation but venues exist that cannot easily be classified into a specific domain, such as Science or Nature. In the MAS data set, papers are labelled as multidisciplinary if they cannot be classified appropriately. These papers are ignored for the experiments conducted in this thesis. More importantly, the size of a domain depends on the publishing sources that are indexed by MAS. Furthermore, the quality of the publishing sources also differ and has an impact on the coverage of the data contained in the MAS data set.

A.2 Evaluation Data Information

Table A.2 shows a list of the conferences and Special Interest Groups for which award papers were selected. The award papers are chosen differently by the various committees but in general the accolade of the award indicates that the papers had a significant impact in their fields. The awards are handed out retrospectively, normally 10 to 12 years after their initial publication. This is shown in the column labelled "Waiting" which contains the number of years that elapse since publication until the awards are handed out. The columns "MAS" and "DBLP" indicate the number of award papers that were matched against corresponding paper entries in the MAS and DBLP data sets, respectively. The column "Years" shows the number of years for which the award papers were collected.

Table A.2: A list of the conferences for which award papers (high-impact papers) were selected.

Venue	Award Name	MAS	DBLP	Years	Waiting	Src.
AAAI	Classic Paper Award	21	0	1999-2013	19 years	[74]
ASE	Most Influential Paper Award	5	4	2010-2013	15 ± 1 years	[75]
ICFP	Most Influential ICFP Paper Award	8	7	2006-2013	10 years	[76]
ICSE	Most Influential Paper Award	25	20	1989-2013	$\approx$ 10 years	[77]
ISCA	Influential ISCA Paper Award	11	10	2003-2013	15 years	[78]
OOPSLA	Most Influential OOPSLA Paper Award	8	5	2006-2012	10 years	[79]
PLDI	Most Influential PLDI Paper Award	14	7	2000-2013	10 years	[80]
POPL	Most Influential POPL Paper Award	11	10	2003-2013	10 years	[81]
SIGEVO	SIGEVO Impact Award	3	0	2011-2013	10 years	[82]
SIGCOMM	Test of Time Paper Award	29	22	2006-2013	usually 10	[83]
SIGMETRICS	Test of Time Award	7	5	2010-2013	$\geq$ 10-12 years	[84]
SIGMOD	Test of Time Award	19	19	1999-2013	10 years	[50]
SIGSOFT	Impact Paper Award	29	25	2008-2013	$\geq$ 10 years	[85]
VLDB	VLDB 10 Years Award	17	17	1995-2012	10 years	[49]
	Total	207	151			

Note that for both "SIGEVO" and "AAAI" no award papers were matched against the DBLP data set. "SIGEVO" is the Special Interest Group on Genetic and Evolutionary Computation and organises the "Genetic and Evolutionary Computation Conference" (GECCO). The papers that won the high-impact prize at "SIGEVO" could not be found in the DBLP data set. A search on the DBLP website reveals that the GECCO conference is in fact indexed by DBLP. "AAAI" stands for the Association for the Advancement of Artificial Intelligence and organises the "AAAI Conference on Artificial Intelligence" conference. The papers that won the "AAAI" high-impact award are found in the DBLP

data set but all have zero citations. The reasons for this are unclear since the method used by Tang *et al.* [6] of obtaining the citation information from the DBLP database is unknown.

Table A.3: Conferences, learned societies or Special Interest Groups that award best paper awards which are used as test data in this thesis.

Venue	Papers	Award Years	Source
AAAI	21	1996-2008	[86]
ACL	14	2001-2009	[86]
ASE	76	1995-2010	[87], [88]
CHI	38	2005-2010	[86]
CIKM	6	2004-2009	[86]
CVPR	11	2000-2010	[86]
FOCS	10	2002-2007	[86]
FSE	19	2002-2009	[86]
ICCV	12	1998-2009	[86]
ICDM	9	2001-2009	[89]
ICML	7	1999-2010	[86]
ICSE	31	2003-2010	[86]
IJCAI	16	1997-2009	[86]
INFOCOM	16	1996-2010	[86]
KDD	12	1997-2008	[86]
LISA	7	2002-2009	[90]
NSDI	6	2004-2010	[86]
OSDI	12	1996-2008	[86]
PLDI	10	1999-2010	[86]
PODS	16	1997-2010	[86]
S&P	3	2008-2010	[86]
SIGCOMM	3	2008-2010	[86]
SIGIR	15	1996-2010	[86]
SIGMETRICS	8	1996-2009	[86]
SIGMOBILE	3	2008-2010	[91]
SIGMOD	13	1996-2010	[86]
SODA	3	2009-2011	[86]
SOSP	22	1997-2009	[86]
STOC	14	2003-2010	[86]
UIST	12	1996-2009	[86]
VLDB	6	1997-2007	[86]
WWW	13	1998-2009	[86]
Total	464		

Table A.3 lists the venues that hand out best paper or distinguished paper awards. The number of papers that were found and matched against database entries in the MAS data set are shown in the column "Papers". The column "Award Years" gives the time spans for which best papers were found and matched against corresponding entries in the MAS dataset.

Similarly, Table A.4 lists the venues that award prizes to authors that made significant and innovative contributions to their fields of research. The number of authors that received the associated awards and were found in the MAS data set are listed under "Authors".

Table A.4: Number of authors who received lifetime achievement or contribution award per venue.

Venue	Authors	Award	Source
AAAI	20	ACM - AAAI Allen Newell Award	[92]
ACL	11	ACL Lifetime Achievement Award	[93]
CHI	15	SIGCHI Lifetime Research Award	[94]
ICCV	4	PAMI Azriel Rosenfeld Lifetime Achievement Award	[95]
ICDM	10	Research Contributions Award	[89]
IJCAI	14	Award for Research Excellence	[96]
ISCA	14	ACM SIGARCH Maurice Wilkes Award	[97]
KDD	13	SIGKDD Innovations Award	[98]
PLDI	24	Programming Languages Achievement Award	[99]
SIGACT	12	Knuth Prize	[100]
SIGCOMM	21	Lifetime Contribution Award	[101]
SIGIR	10	Gerard Salton Award	[102]
SIGMETRICS	11	Achievement Award	[103]
SIGMOBILE	14	Outstanding Contributions Award	[104]
SIGMOD	22	SIGMOD Edgar F. Codd Innovations Award	[50]
SIGOPS	14	Mark Weiser Award	[50]
SIGSIM	6	ACM SIGSIM Distinguished Contributions Award	[105]
SIGSOFT	23	ACM SIGSOFT Outstanding Research Award	[106]
USENIX	10	USENIX Lifetime Achievement Award	[107]
Total	268		

A.3 Additional Results

Tables A.5 through A.8 show the ranking results of PageRank, NewRank, YetRank and SceasRank by listing the 10 highest ranked papers. In each table, the citation counts, the publication years and the average age of the citations are given for each paper.

Table A.5: Top 10 highest ranked papers according to PageRank with the default damping factor of $\alpha = 0.85$.

Score	Title	Cites	Year	Cite Age
0.00069582	MODELTEST: testing the model of DNA substitution	8234	1998	8.82
0.00065912	MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment	5875	2004	3.99
0.00065874	Applied Regression Analysis	4101	1968	30.26
0.00061662	Matrix Computations	7822	1986	17.06
0.00061064	MRBAYES: Bayesian inference of phylogenetic trees	4317	2001	6.23
0.00055178	A mathematical theory of communication	5602	2001	0.63
0.00054196	MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers	284	1994	10.40
0.00052079	Optimization by Simulated Annealing	5872	1983	19.94
0.00049365	Fuzzy Sets	8954	1965	38.71
0.00048554	The rapid generation of mutation data matrices from protein sequences	1198	1992	13.75
		5225.90	1989.02	14.98

It should be noted that the paper with the title "A mathematical theory of communication" in Tables A.5 and A.6 has a wrong year associated with it and was actually published in 1964 and therefore the average citation age is skewed.

Table A.9 shows the 10 highest ranked authors according to the Author-Level Eigenfactor method computed on the MAS CS citation network. The ranks that the authors obtained according to their citation counts, the h-index, the g-index and the *i10*-index are also listed in this table.

Table A.6: Top 10 highest ranked papers according to NewRank with the default parameters $\alpha = 0.85$ and $\tau = 4.0$.

Score	Title	Cites	Year	Cite Age
0.00102547	MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment	5875	2004	3.99
0.00094561	MRBAYES: Bayesian inference of phylogenetic trees	4317	2001	6.23
0.00076333	A mathematical theory of communication	5602	2001	0.63
0.00072622	Analysis of Variance for Gene Expression Microarray Data	547	2000	5.55
0.00066662	Haploview: analysis and visualization of LD and haplotype maps	3275	2005	3.62
0.00066618	On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data	196	1999	6.10
0.00063196	MrBayes 3: Bayesian phylogenetic inference under mixed models	4660	2003	5.43
0.00057578	MODELTEST: testing the model of DNA substitution	8234	1998	8.82
0.00056433	MEGA2: molecular evolutionary genetics analysis software	3339	2001	4.67
0.00053589	MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers	284	1994	10.40
		3632.90	2000.60	6.12

Table A.10 shows the complete list of venues that award best paper prizes at conferences. Best paper awards are usually awarded to more than one paper in a year at a conference. The results in this table consider all papers that won the best paper awards. The column "Entire Network" displays the precision (MAP@10) of the venues in predicting high-impact papers if the entire MAS Computer Science network is used. Alternatively, "Subset Network" shows the precision values if the input graph is truncated to 5 years after the papers won the best-paper awards.

Table A.11 shows the precision of the award committees in predicting high-impact papers based on a single best paper per year at the associated conferences. The best paper with the highest citation count was chosen for the test data. The column "MAP@10" displays the MAP values with the input network truncated to 5 years after the papers won the best-paper awards.

Table A.7: Top 10 highest ranked papers according to YetRank with the default parameters $\alpha = 0.85$ and $\tau = 4.0$. The target and window sizes, used by the Impact Factor method in YetRank, were set to 5 and 1 years, respectively.

Score	Title	Cites	Year	Cite Age
0.00063355	A method for obtaining digital signatures and public-key cryptosystems	3223	1978	23.79
0.00048593	Matrix Computations.	7822	1986	17.06
0.00042879	Digital communications	3961	1985	20.36
0.00039495	Optimization by Simulated Annealing	5872	1983	19.94
0.00036916	Congestion avoidance and control	1876	1988	14.03
0.00035976	Fuzzy Sets	8954	1965	38.71
0.00035348	Support-vector networks	2335	1995	11.27
0.00030729	A tutorial on hidden Markov models and selected applications in speech recognition	3590	1989	16.16
0.00029949	Learning internal representations by error propagation	2678	1986	14.00
0.00029015	Induction of decision trees	2973	1986	15.79
		4328.40	1984.10	19.15

Table A.8: Top 10 highest ranked papers according to SceasRank with the default parameters $\alpha = 0.85$, $a = e$ and $b = 1$.

Score	Title	Cites	Year	Cite Age
0.00173798	MODELTEST: testing the model of DNA substitution	8234	1998	8.82
0.00164546	MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment	5875	2004	3.99
0.00127838	Applied Regression Analysis	4101	1968	30.26
0.00118639	Working memory	3949	2010	-6.97
0.00111718	A mathematical theory of communication	5602	2001	0.63
0.00105675	Matrix Computations	7822	1986	17.06
0.00105666	MEGA2: molecular evolutionary genetics analysis software	3339	2001	4.67
0.00105257	MRBAYES: Bayesian inference of phylogenetic trees	4317	2001	6.23
0.00096612	An Algorithm for Least-Squares Estimation of Nonlinear Parameters	3624	1963	37.26
0.00095463	Haploview: analysis and visualization of LD and haplotype maps	3275	2005	3.62
		5013.80	1993.70	11.92

Table A.9: The top 10 authors according to the Author-Level Eigenfactor method and their corresponding ranks using citation counts with self-citations (CCRS), the *h*-index, the *g*-index and the *i10*-index.

Name	**Cites**	**Papers**	**Avg. Cites**	**CCRS**	***h***	***g***	***i10***
Charles A.T. Hoare	8 987	153	58.74	69	192	49	345
Robert E. Tarjan	21 414	315	67.98	2	1	2	3
Edsger W. Dijkstra	4 973	68	73.13	310	1 357	267	2 778
Donald E. Knuth	5 871	129	45.51	208	299	154	412
Leslie Lamport	15 321	167	91.74	10	52	8	133
Edgar F. Codd	5 038	42	119.95	299	3 029	1 517	3 579
Lofti A. Zadeh	16 481	92	179.14	6	369	57	1 109
Ronald L. Rivest	13 811	178	77.59	18	40	15	121
John McCarthy	4 157	71	58.55	350	1 269	265	2 348
Leslie Valiant	4 673	107	43.67	58	127	42	316
Average	10 596.8	132.2	80.16	133.0	673.5	237.6	1114.4

Table A.10: The precision of the award committees in identifying high-impact papers based on the papers that won best-paper awards at the associated conferences.

	Entire Network			Subset Network		
Conference	**Count**	**In-Degree**	**MAP@10**	**Count**	**In-Degree**	**MAP@10**
SOSP	22	118.36	0.595	19	66.89	0.577
OSDI	12	117.75	0.549	12	78.42	0.544
SIGMETRICS	8	67.75	0.486	7	54.71	0.525
FOCS	10	65.60	0.463	10	57.90	0.495
ACL	14	72.50	0.467	11	68.00	0.457
FSE	19	44.68	0.363	17	41.29	0.414
UIST	12	48.92	0.420	11	34.82	0.384
CIKM	6	22.33	0.274	5	25.80	0.329
CVPR	11	157.55	0.243	9	126.22	0.307
ICCV	12	121.42	0.175	11	77.82	0.297
PLDI	10	80.40	0.228	8	65.88	0.295
STOC	14	47.71	0.320	10	54.00	0.286
KDD	12	117.58	0.369	12	69.00	0.281
ICDM	9	30.67	0.281	8	28.00	0.271
VLDB	6	89.67	0.250	6	73.67	0.250
SIGCOMM	3	23.00	0.187	1	49.00	0.250
NSDI	6	40.33	0.270	3	70.00	0.242
ASE	76	19.87	0.261	73	14.05	0.240
ICSE	31	38.97	0.172	22	46.09	0.227
WWW	13	132.08	0.199	12	88.00	0.218
LISA	7	8.00	0.264	6	8.00	0.217
SIGMOD	13	111.38	0.173	11	75.36	0.195
INFOCOM	16	100.13	0.166	14	74.71	0.189
AAAI	21	50.76	0.153	21	30.48	0.172
PODS	16	60.31	0.209	14	38.71	0.150
ICML	7	41.71	0.139	5	47.60	0.150
CHI	38	18.45	0.106	20	28.75	0.080
IJCAI	16	25.63	0.007	14	19.57	0.009
S&P	3	15.33	0.245	1	25.00	0.125
SIGIR	15	63.80	0.133	13	40.85	0.089
SIGMOBILE	3	6.33	0.208	1	11.00	0.000
SODA	3	7.33	0.050	0	0.00	0.000
	464	61.45	0.263	387	49.68	0.258

Table A.11: Precision of award committees in predicting high-impact papers based on the single papers that won a best-paper award with the highest citation count for each year in which the best-paper prize was awarded.

Conference	Nr. Years	In-Degree	Avg. Papers	MAP@10
SOSP	7	106.50	2.71	0.639
SIGMETRICS	6	51.50	1.17	0.483
ACL	8	78.50	1.38	0.458
FOCS	6	68.33	1.67	0.410
FSE	7	71.57	2.43	0.405
OSDI	6	102.43	2.00	0.392
UIST	10	34.90	1.10	0.375
ICCV	8	112.17	1.38	0.344
CIKM	5	25.80	1.00	0.329
AAAI	11	49.64	1.91	0.318
CVPR	8	138.63	1.13	0.286
KDD	12	69.00	1.00	0.281
PLDI	6	83.67	1.33	0.274
ICDM	8	28.00	1.00	0.271
SIGCOMM	1	49.00	1.00	0.250
VLDB	6	73.67	1.00	0.250
STOC	6	70.83	1.67	0.249
NSDI	3	70.00	1.00	0.242
LISA	5	11.00	1.20	0.217
WWW	11	84.45	1.09	0.211
SIGMOD	11	75.36	1.00	0.195
INFOCOM	11	85.67	1.27	0.155
ICML	5	47.60	1.00	0.150
PODS	12	43.33	1.17	0.142
S&P	1	25.00	1.00	0.125
ASE	13	39.79	5.62	0.119
SIGIR	13	40.85	1.00	0.089
CHI	4	52.00	5.00	0.083
ICSE	6	80.00	3.67	0.082
IJCAI	6	28.67	2.33	0.000
SIGMOBILE	1	11.00	1.00	0.000
SODA	0	0.00	0.00	0.000
	223	59.65	1.61	0.245